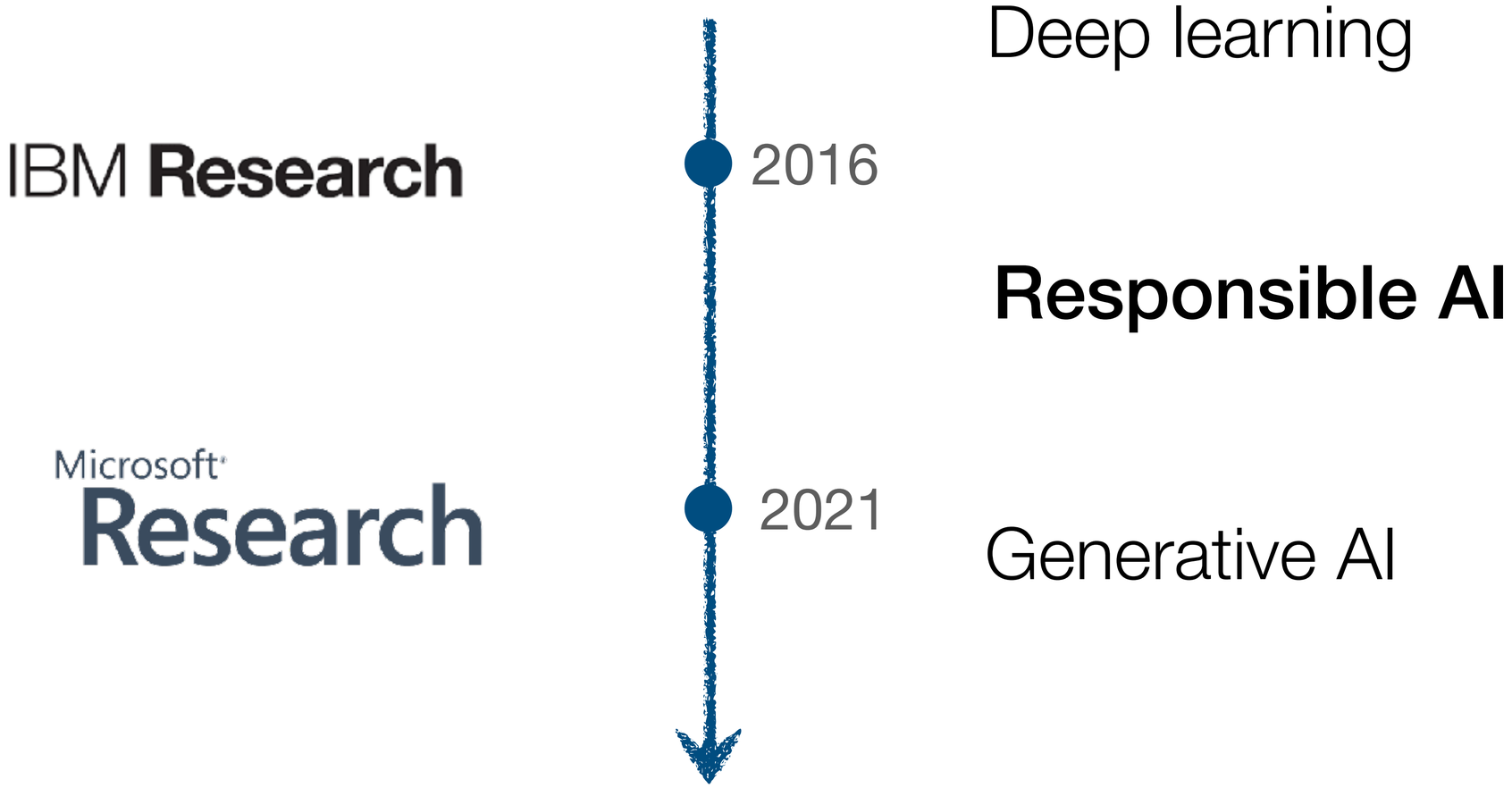
Design for Responsible AI (RAI)

Q. Vera Liao

Microsoft Research

University of Michigan

My Journey into *Studying* UX Design Practices in the Age of AI



Learn from UX practitioners

How might current AI technologies fail to meet people's needs?

Empower UX practitioners

What role can and should design play in making AI technology "good" for people?

How to better support designers to play such a role through a new design toolboxe (tools, guidances, organizational practices, etc.)

Learn from UX practitioners

How might current AI technologies fail to meet people's needs?

Empower UX practitioners

What role can and should design play in making AI technology "good" for people?

How to better support designers to play such a role through a new design toolboxe (tools, guidances, organizational practices, etc.)

AI as new design materials

What are the challenges?

Challenges Working with AI as Design Materials

Output complexity

Capability complexity

Materialistic uncertainty

(Yang et al., *2020;* Subramonyam et al., 2021 *)*

Challenges Working with AI as Design Materials

Output complexity	Challenge with understanding the material
Capability complexity	Challenge with choosing (or having) the right material
Materialistic uncertainty	Challenge with mutual shaping of design and material

NEWSLETTERS · EYE ON A.I.

ChatGPT's inaccuracies are causing real harm

TECH • ARTIFICIAL INTELLIGENCE

The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter

AI Bias In Recruitment: Ethical Implications And Transparency

AP

AI chatbots are supposed to improve health care. But research says some are perpetuating racism

Snapchat's AI chatbot may pose privacy risk to children, says UK watchdog

Ethical concerns mount as AI takes bigger decision-making role in more industries

OP-ED CONTRIBUTOR

When an Algorithm Helps Send You to Prison

WHO warns against bias, misinformation in using AI in healthcare

How Artificial Intelligence Can Deepen Racial and Economic Inequities

Here are dozens of ways AI could be used for harm — and some too scary to test

NEWSLETTERS · EYE ON A.I.

ChatGPT's inaccuracies are causing real harm

Ethical concerns mount as AI takes bigger decision-making role in more industries

TECH • ARTIFICIAL INTELLIGENCE

The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter

OP-ED CONTRIBUTOR

When an Algorithm Helps Send You to Prison

Responsible AI (RAI): Develop and deploy AI technologies responsibly by mitigating harms to people, community and society

AP

AI chatbots are supposed to improve health care. But research says some are perpetuating racism

Deepen Racial and Economic Inequities

Snapchat's AI chatbot may pose privacy risk to children, says UK watchdog

Here are dozens of ways AI could be used for harm — and some too scary to test

RAI: Sociotechnical Harm/Risk Management

"AI systems are inherently socio-technical in nature, meaning they are influenced by societal dynamics and human behavior. AI risks —and benefits—can emerge from the ***interplay of technical aspects combined with societal factors*** *related to how a system is used, its interactions with other systems, who operates it, and the context in which it is deployed."*

NIST (US National Institute of Standards and Technology) AI Risk Management Framework

AI ___ creates ___ **harm** in the *context of* ___

AI bias creates ___ **harm** in the *context of* ___

AI bias creates **allocation harm** in the *context of hiring tools*

AI bias creates **representation harm** in the *context of image generation tools*

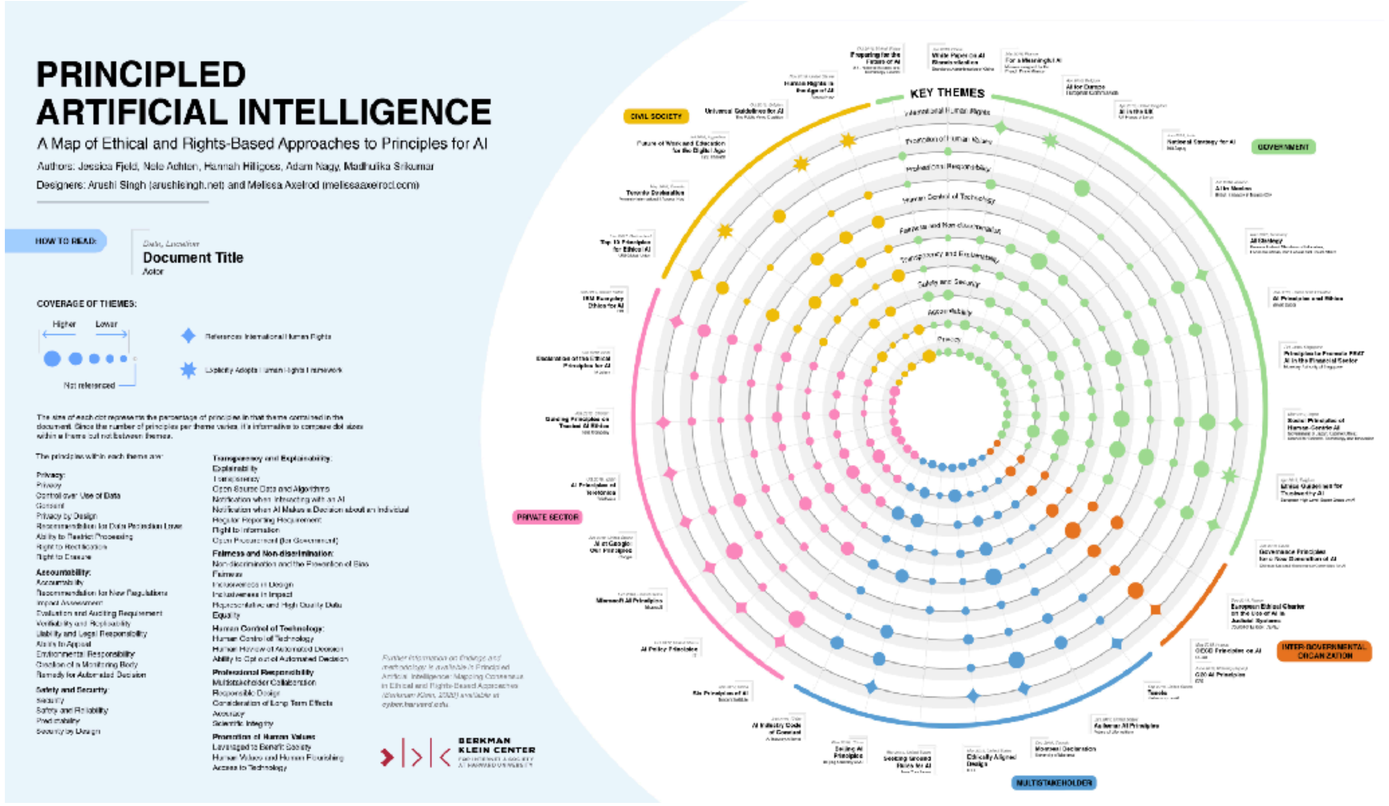
AI bias creates **quality-of-service harm** in the *context of facial recognition (as a service) tools*

AI opaqueness creates

harms of

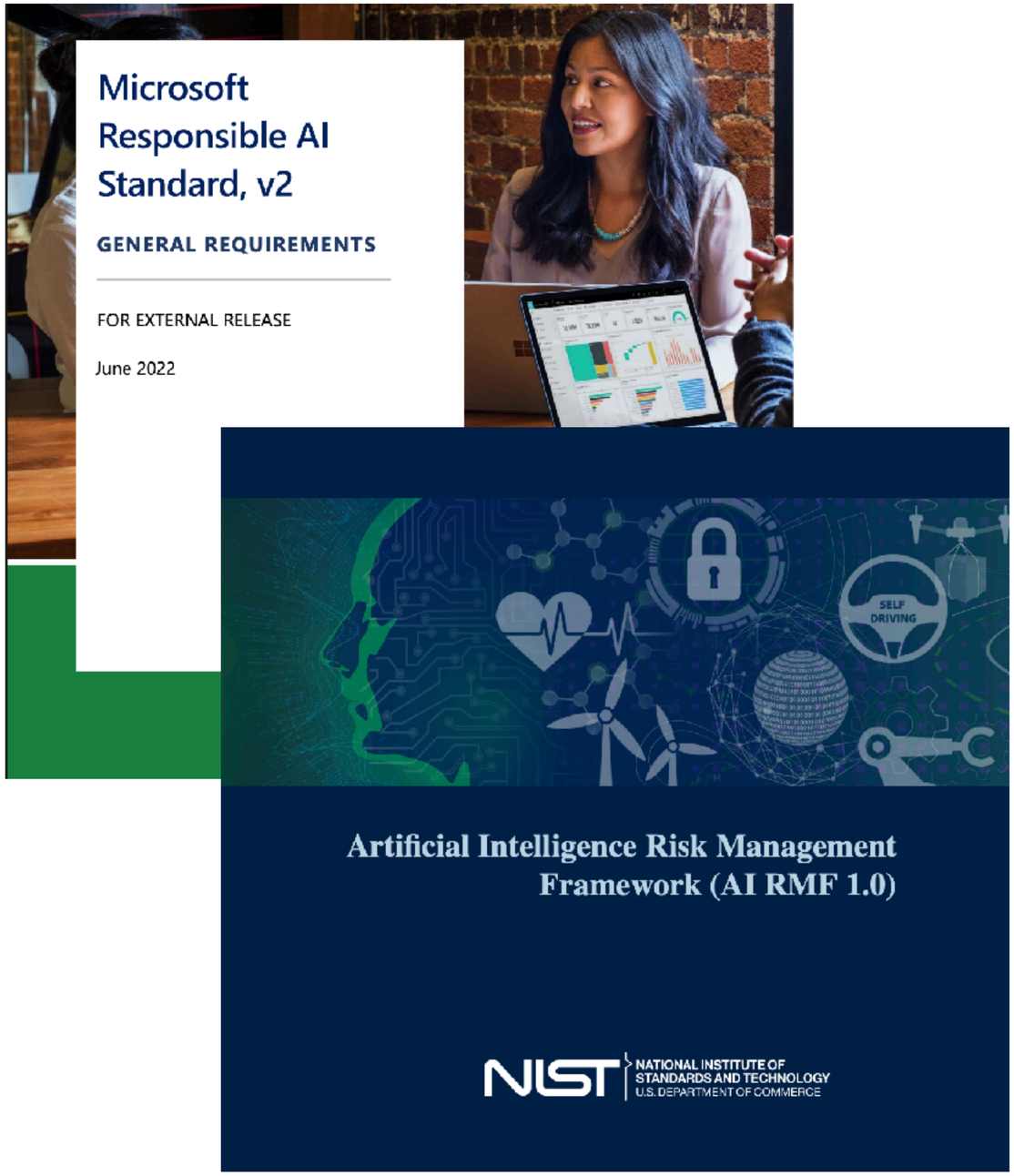
- **mistrust**
- **misuse**
- **loss of agency**
- **deprivation of recourse**
- **...**

RAI: Principle Based Approaches



Mapping of 36 ethical and responsible AI frameworks (Berkman Klein Center)

- Fairness and Non-discrimination
- Transparency and Explainability
- Human Control
- Safety and Security
- Accountability
- Privacy
- Promotion of Human Values



How is **Responsible AI development** conceived in RAI policy?

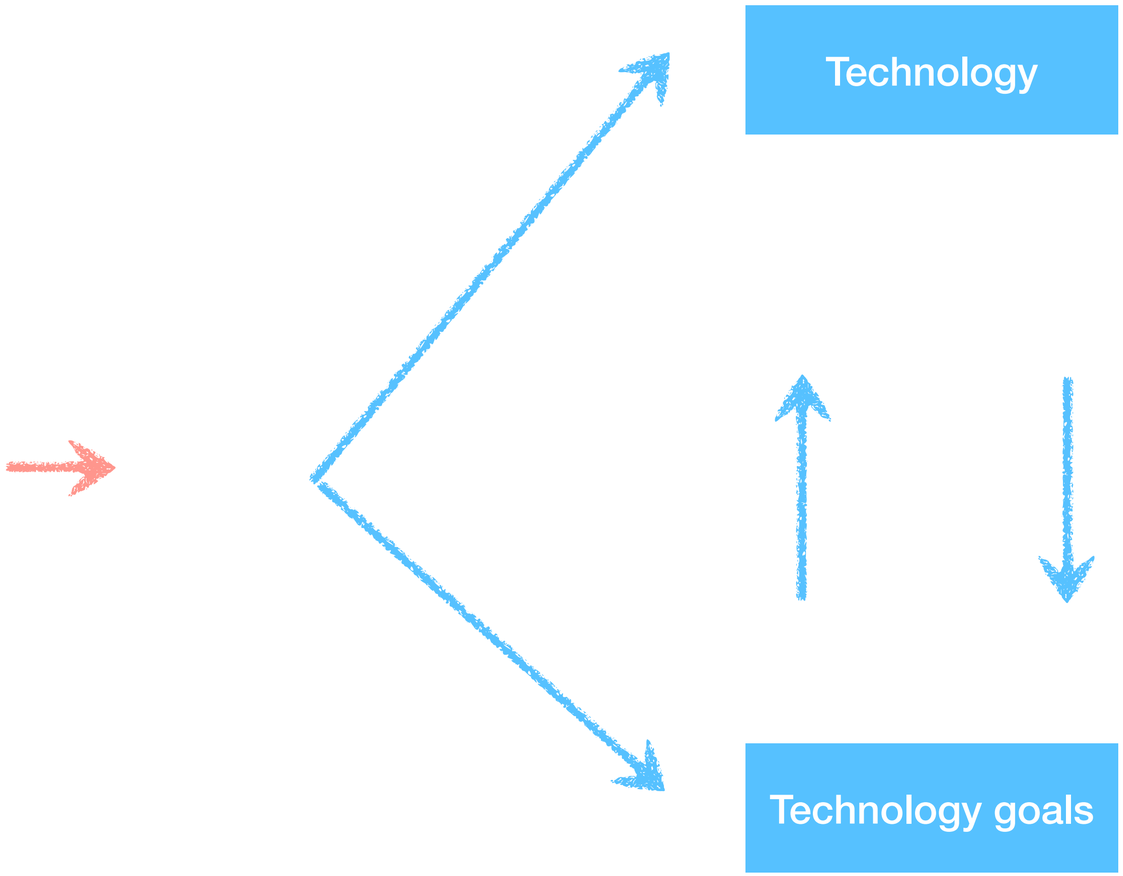
A simplified view…

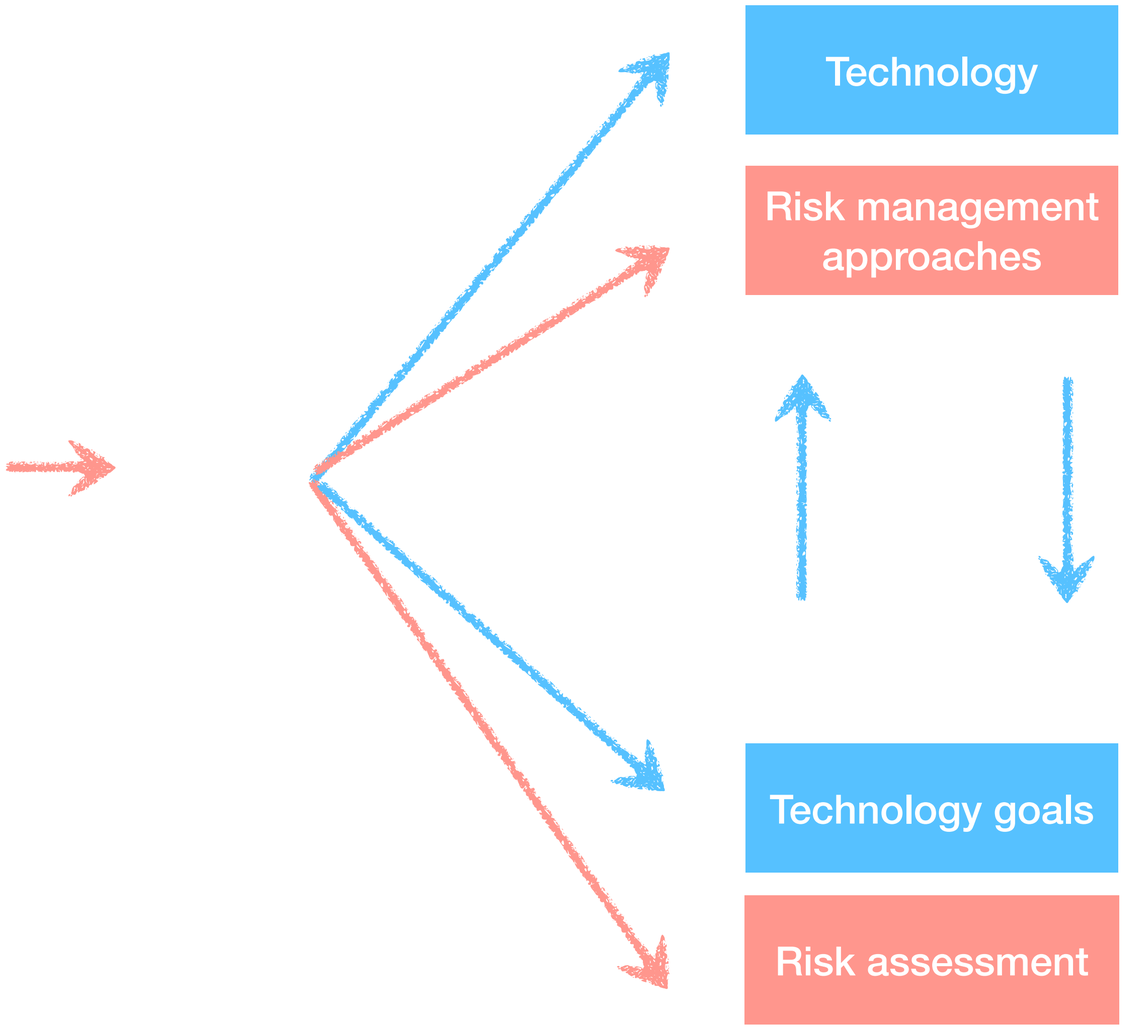
Technology

Technology goals

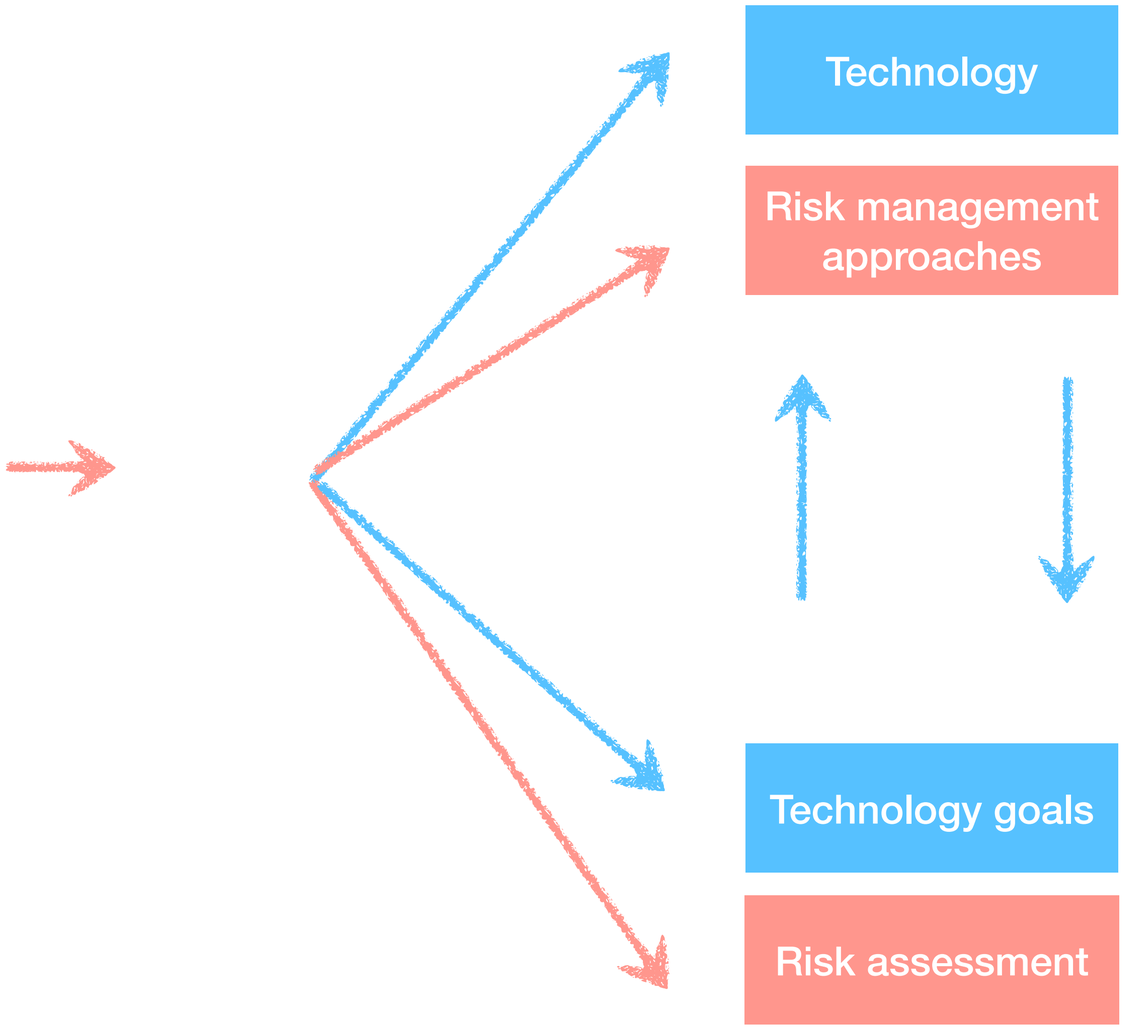
Technology

Technology goals


Technology
Technology goals


Technology
Risk management approaches
Technology goals
Risk assessment

- Fairness and Non-discrimination
- Transparency and Explainability
- Human Control
- Safety and Security
- Accountability
- Privacy
- Promotion of Human Values



Shared **sociotechnical perspective** in **RAI** and **UX** practices

Research Thread 1: Empower Designers in Responsible AI Development

Envisioning the sociotechnical and managing AI risks

Fairness and Non-discrimination

Transparency and Explainability

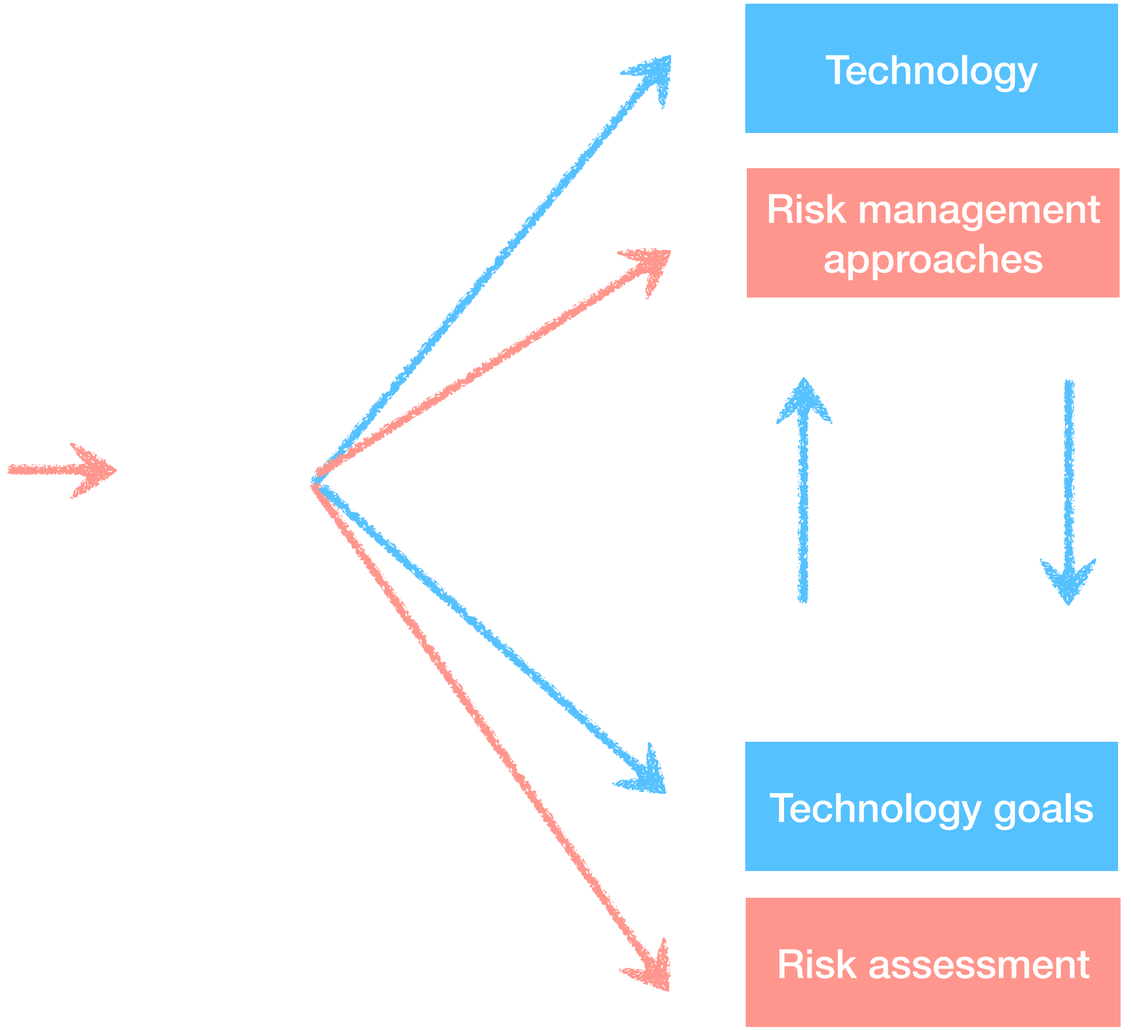
Human Control

Safety and Security

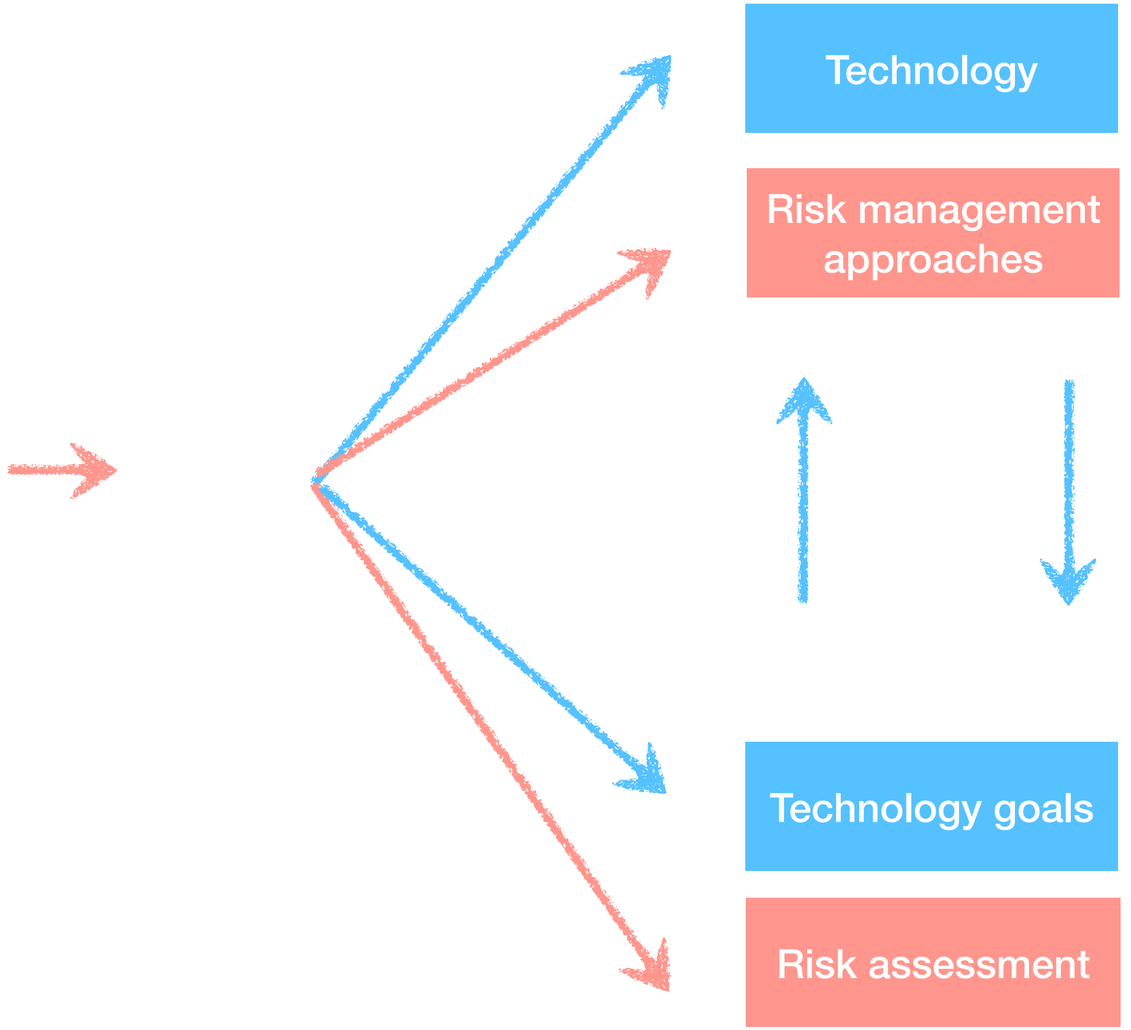
Accountability

Privacy

Promotion of Human Values

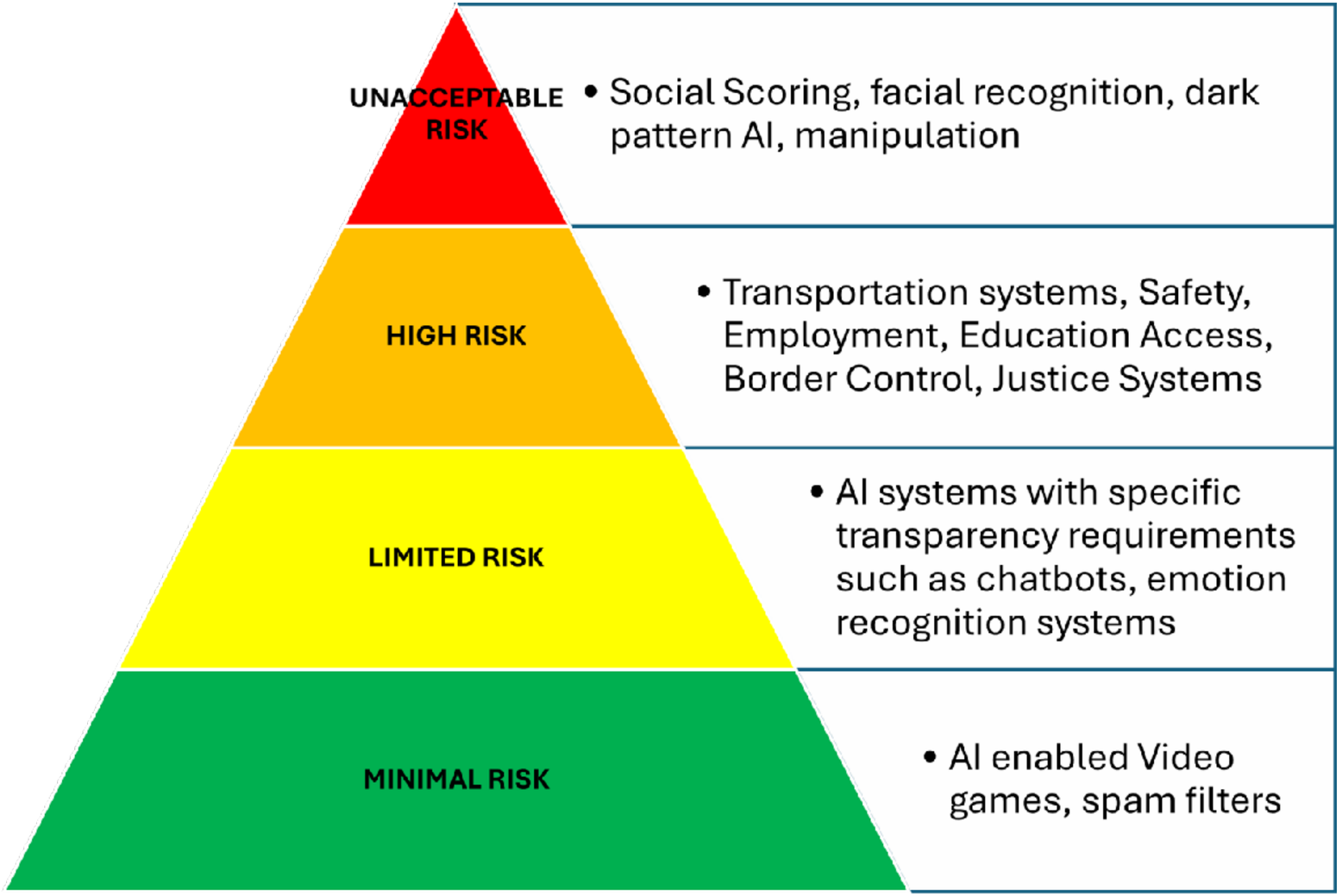


Transparency and Explainability



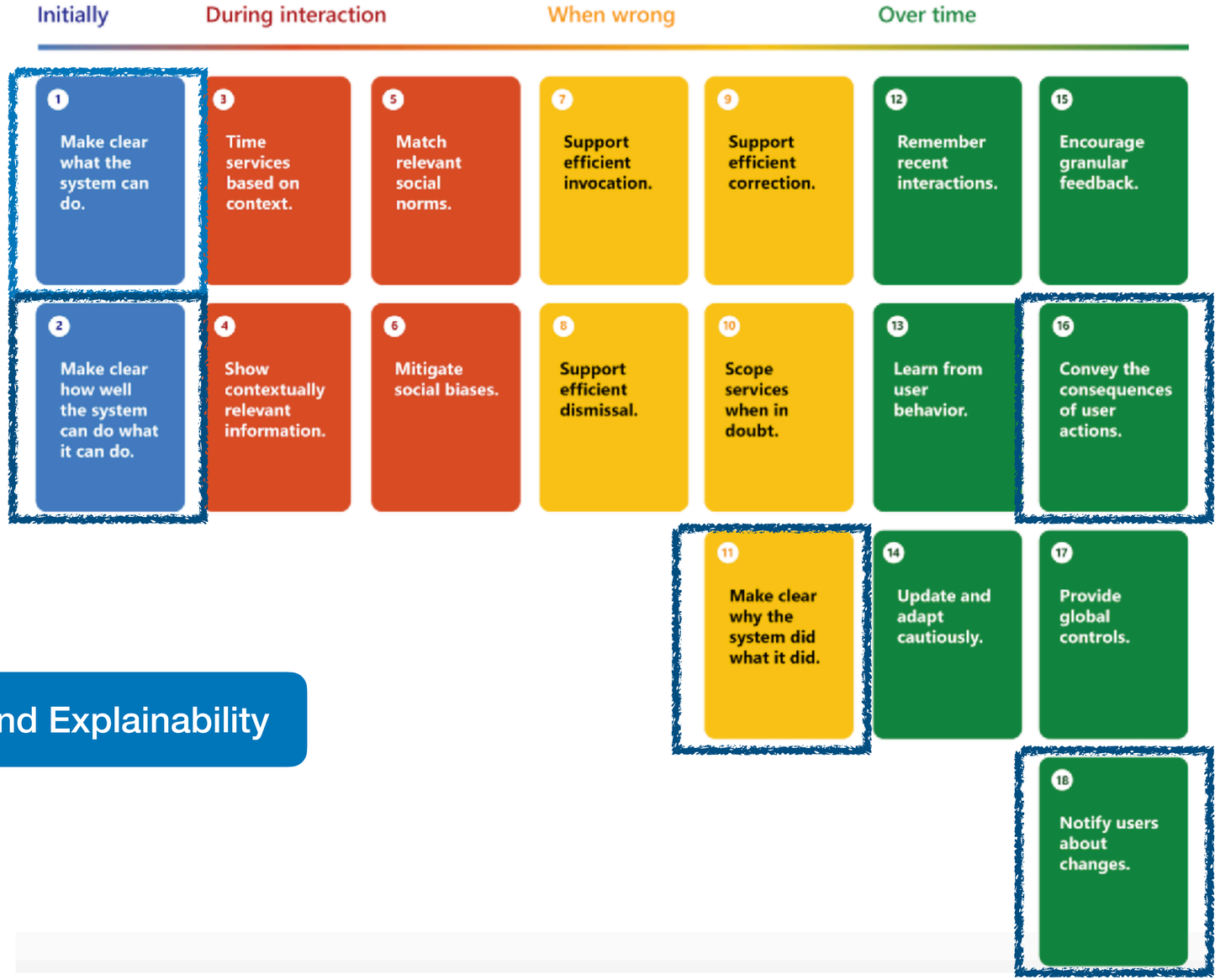
Transparency and Explainability

Risk management approaches


UNACCEPTABLE RISK
• Social Scoring, facial recognition, dark pattern AI, manipulation
HIGH RISK
• Transportation systems, Safety, Employment, Education Access, Border Control, Justice Systems
LIMITED RISK
• AI systems with specific transparency requirements such as chatbots, emotion recognition systems
MINIMAL RISK
• AI enabled Video games, spam filters

EU AI Act

Design Guidelines for Human-AI Interaction (HAX)



Transparency and Explainability

(Amershi et al., *2019)*

Transparency and Explainability

Risk management approaches

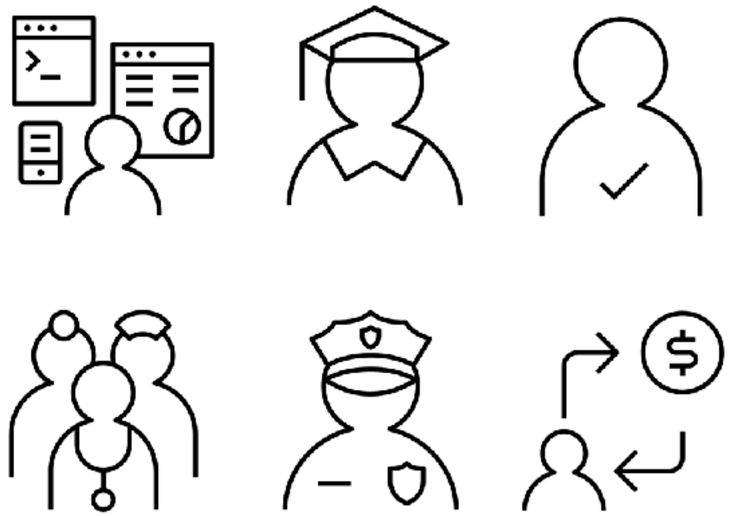
Learn from UX practitioners:

How to provide AI explanations to promote human understanding and oversight?

Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *CHI 2020*

> *It remains in this weird limbo where people know it [explainability] is important. People see it happen. They don't know how to make it happen. And everybody's feeling their way in the dark with no lights.*
>
> (P6)

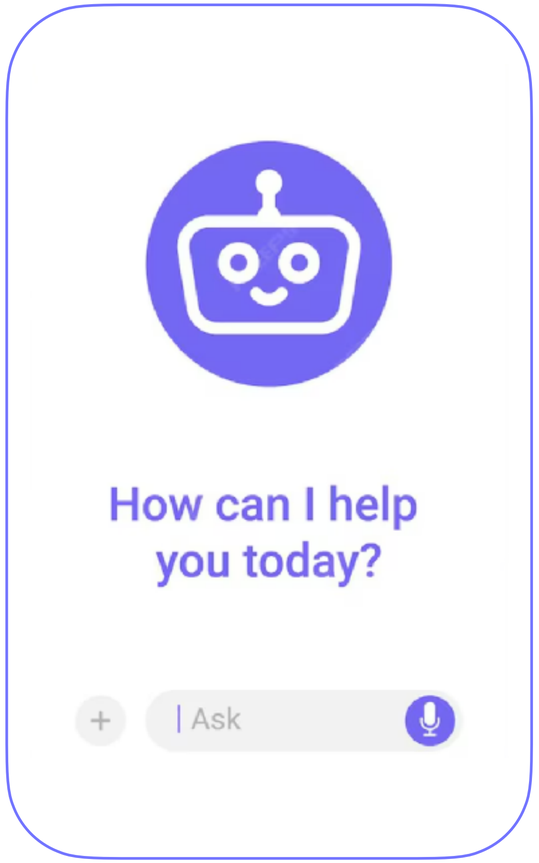
Design Challenge 1: Variability of Users' Explainability Needs



Diverse reasons for AI explanations

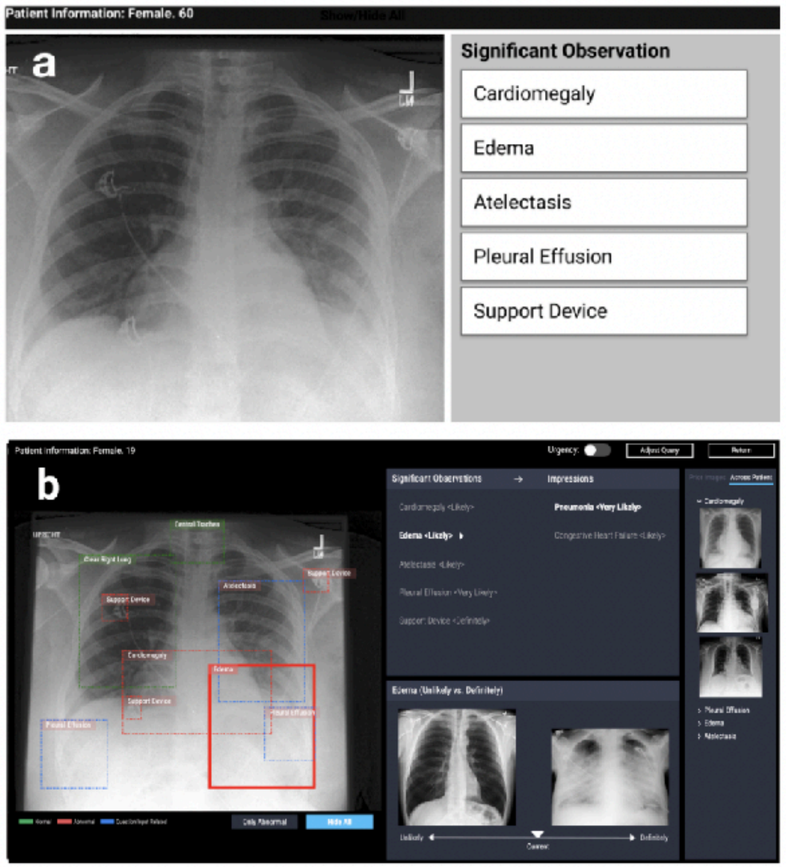
- To assess AI capabilities
- To improve AI capabilities
- To gain further insights for decision-making
- To adapt interaction
- To avoid rights violations
- To seek recourse
- …

Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *CHI 2020*



What kinds of questions can it answer?

How does the bot provide answers? What data does it have access to?



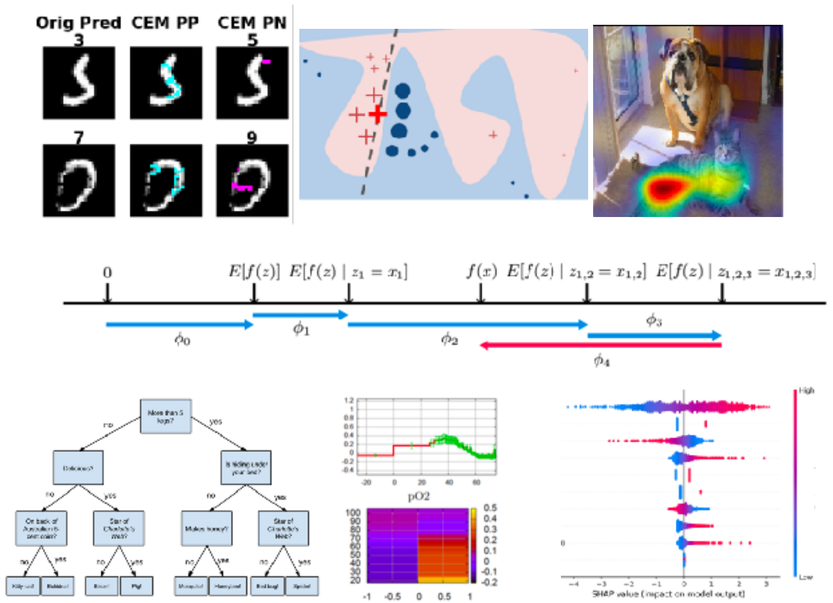
CheXplain

(Xie et al, 2020)

Why is the AI system giving this diagnosis?

Why is the AI system not giving a diagnosis that I would expect?

Design Challenge 2: Complexity of Explainable AI (XAI) Algorithms



- Dozens of algorithms in XAI data science toolkits (hundreds more from research)
- Difficulty in understanding and having the right material as UX practitioners are often not involved in the choosing
- Gaps between algorithmic output and design intent

Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *CHI 2020*

XAI algorithms

> *…finding the right pairing to put the ideas of what's right for the user together with what's doable given the tools or the algorithms"* (P8)

User's explainability needs

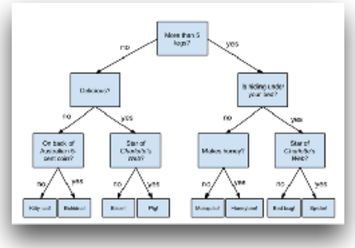
XAI algorithms

> *…finding the right pairing to put the ideas of what's right for the user together with what's doable given the tools or the algorithms"* (P8)

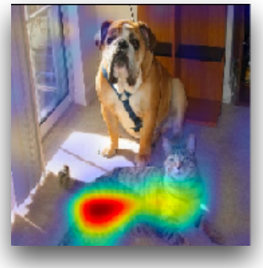
User's explainability needs

How can (envisioned) user needs and interactions drive the choices of XAI algorithms as design material?

Sociotechnical Abstraction: Questions Answerable by XAI Algorithms



Global explanation: *How* does the AI make predictions?



Feature-importance explanation: *Why* is the AI giving this prediction?

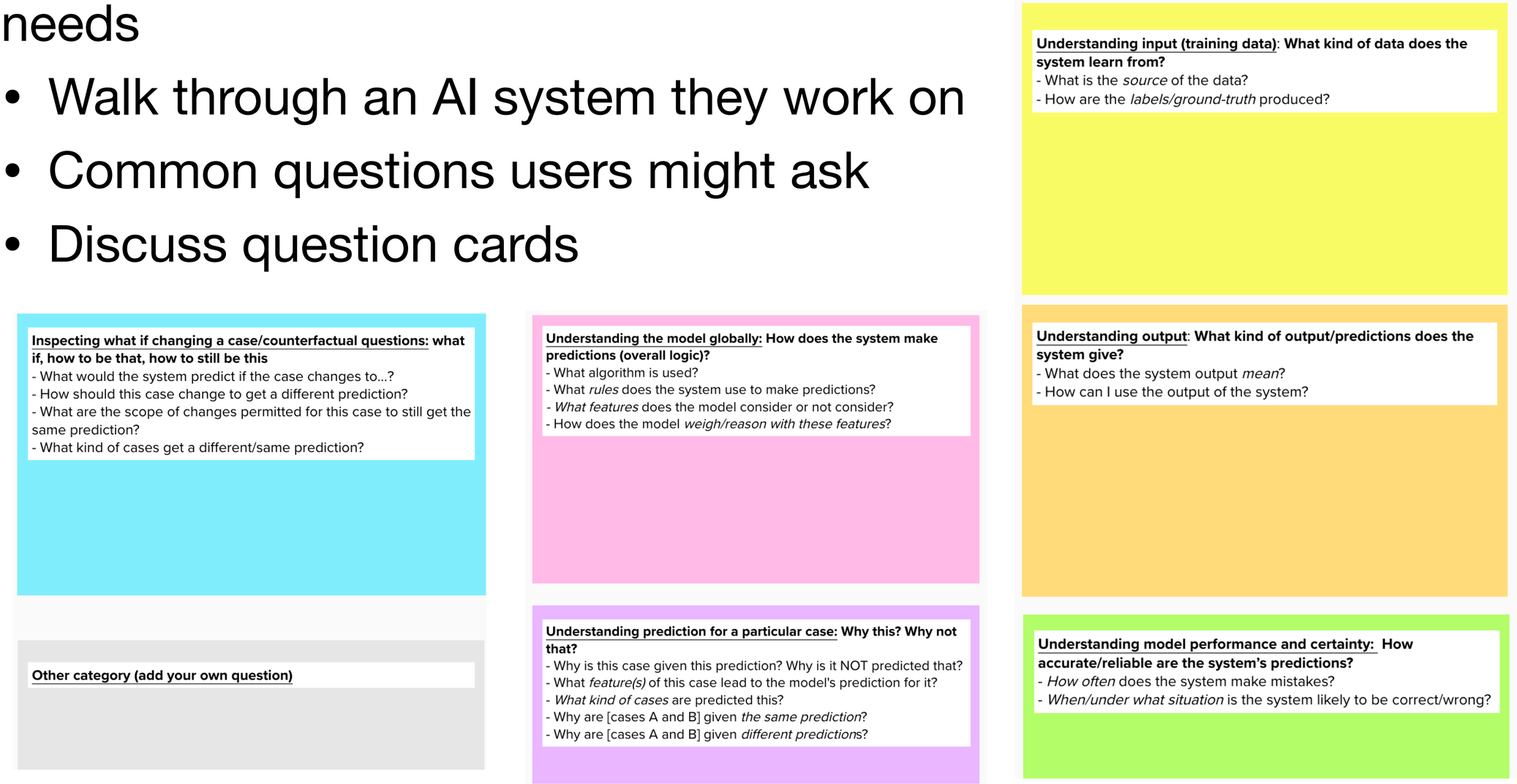
Classified as “9”

Classified as “4”

Counterfactual explanation: *How to* get a *different* prediction?

Engage with UX practitioners to map the space of user questions as explainability needs

- Walk through an AI system they work on
- Common questions users might ask
- Discuss question cards



Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *CHI 2020*

XAI Question Bank

Data

- **What kind of data was the system trained on?**
- What is the source of the training data?
- How were the labels/ground-truth produced?
- What is the sample size of the training data?
- What dataset(s) is the system NOT using?
- What are the potential limitations/biases of the data?
- What is the size, proportion, or distribution of the training data with given feature(s)/feature-value(s)?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- What is the scope of the system's capability? Can it do…?
- How is the output used for other system component(s) ?
- How should I best utilize the output of the system?
- How should the output fit in my workflow?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/ incorrect?
- What are the limitations of the system?
- What kind of mistakes is the system likely to make?
- Is the system's performance good enough for…?

How
(global model-wide explanation)

- **How does the system make predictions?**
- What features does the system consider?
  - Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What kind of rules does it follow?
  - How does [feature X] impact its predictions?
  - What are the top rules/features that determine its predictions?
- What kind of algorithm is used?
  - How were the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance determine the system's prediction of it?
- Why are [instance A and B] given the same prediction?

Why not

- **Why is this instance NOT predicted to be [a different outcome Q]?**
- Why is this instance predicted [P instead of a different outcome Q]?
- Why are [instance A and B] given different predictions?

How to be that
(a different prediction)

- **How should this instance change to get a different prediction Q?**
- What is the minimum change required for this instance to get a different prediction Q?
- How should a given feature change for this instance to get a different prediction Q?
- What kind of instance is predicted of [a different outcome Q]?

How to still be this
(the current prediction)

- **What is the scope of change permitted for this instance to still get the same prediction?**
- What is the range of value permitted for a given feature for this prediction to stay the same?
- What is the necessary feature(s)/feature-value(s) present or absent to guarantee this prediction?
- What kind of instance gets the same prediction?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if a given feature changes to…?
- What would the system predict for [a different instance]?

Others

- How/why will the system change/adapt/improve/drift over time? (change)
- Can I, and if so, how do I, improve the system? (improvement)
- Why is the system using or not using a given algorithm/feature/rule/dataset? (follow-up)
- What does [a machine learning terminology] mean? (terminological)
- What are the results of other people using the system? (social)

Map Questions to XAI Approaches

Question	Explanations	Example XAI techniques
Global how (global model-wide)	• Describe the general model logic as feature impact*, rules✣ or decision-trees• (sometimes need to explain with a surrogate simple model) • If the user is only interested in a high-level view, describe what are the top features or rules considered	ProfWeight*✣•, Global Feature Importance*, PDP*, DT Surrogate•
Why	• Describe how features of the instance, or what key features, determine the model's prediction of it* • Or describe rules✣ that the instance fits to guarantee the prediction✣ • Or show similar examples• with the same predicted outcome to justify the model's prediction	LIME*, SHAP*, LOCO*, Anchors✣, ProtoDash•
Why not (a different prediction)	• Describe what features of the instance determine the current prediction and/or with what changes the instance would get the alternative prediction* • Or show prototypical examples✣ that had the alternative outcome	CEM* , Counterfactuals✣ , ProtoDash✣ (on alternative prediction)
How to be that (a different prediction)	• Highlight feature(s) that if changed (increased, decreased, absent, or present) could alter the prediction to the alternative outcome, often with minimum effort required* • Or show examples with minimum differences but had the alternative outcome✣	CEM*, Counterfactuals✣, DiCE✣
How to still be this (the current prediction)	• Describe features/feature ranges* or rules✣ that could guarantee the same prediction • Show examples that are different from the particular instance but still had the same outcome	CEM*, Anchors✣
What if	• Show how the prediction changes corresponding to the inquired change of input	PDP, ALE
Performance	• Provide performance metrics of the model • Show uncertainty information for each prediction • Describe potential strengths and limitations of the model	Precision, Recall, Accuracy, F1, AUC Uncertainty Qauntification 360 FactSheets, Model Cards
Data	• Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.	FactSheets, DataSheets
Output	• Describe the scope of output or system functions. • Suggest how the output should be used for downstream tasks or user workflow	FactSheets, Model Cards

Question-Driven XAI Design

Step 1

Collect user questions

Elicit user needs for explainability as questions

Understand user intention and expectation of asking these questions

Optionally, Question Bank could be used as a checklist to help guide the elicitation

Step 2

Analyze and identify key user questions and requirements

Cluster similar questions and prioritize with your team the categories to focus on

Question Bank suggests 9 categories for supervised ML. But the categorization could be flexible

Identify key user requirements by analyzing users' intention and expectation to ask the questions

Step 3

Map questions to modeling solutions

Map prioritized question categories to candidate explainability solutions

Sometimes it requires working with data scientists to identify the right algorithms to generate explanations

A mapping guide for supervised ML is provided for reference

Step 4

Iteratively **design** and **evaluate**

Evaluate the candidate design with the key user requirements, iteratively improve the design and modeling solution to close the gaps

It would be ideal to get feedback from actual users in these iterations

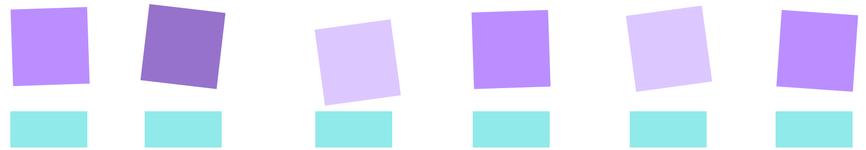
Liao et al. Question-Driven Design Process for Explainable AI User Experiences. arXiv

Problem description	An AI based dashboard presents patients' **readmission risk scores** to help clinicians to identify high-risk patients and appropriate interventions at discharge time

Step 1: **Collect** user questions

Identify what aspects of AI needs to be explained by eliciting user questions

Also gather user intention and expectation of asking these user questions

Problem description	An AI based dashboard presents patients' **readmission risk scores** to help clinicians to identify high-risk patients and appropriate interventions at discharge time
Tasks involved (optional)	
Questions from User 1	
Questions from User 2	

Step 2: **Analyze** to identify key user questions and user requirements

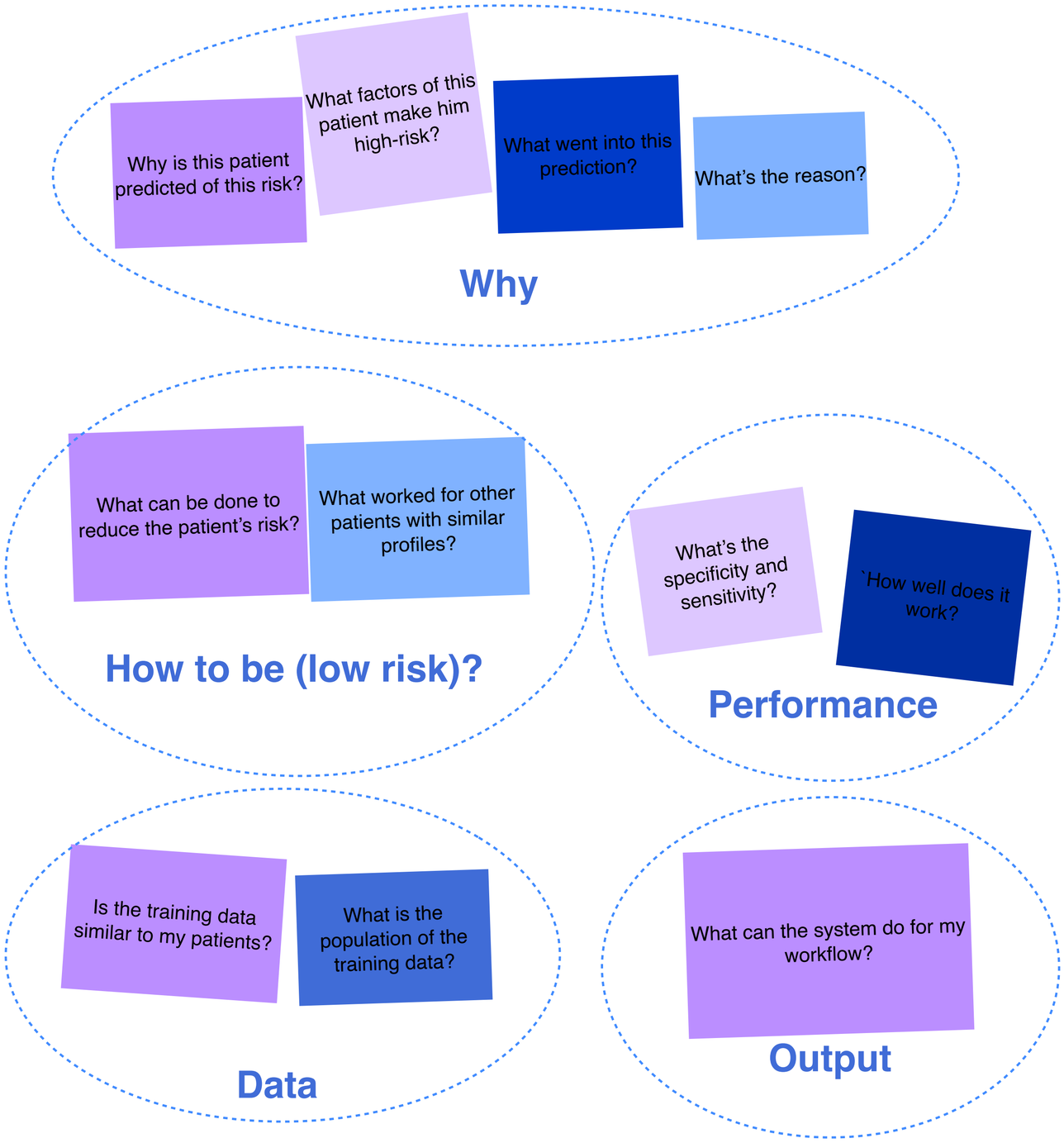
---

Cluster similar questions across users into key categories.

Question Bank could guide the analysis.

Identify which categories should the team focus on

Analyze user comments on intention and expectation of asking the questions to identify key user requirements



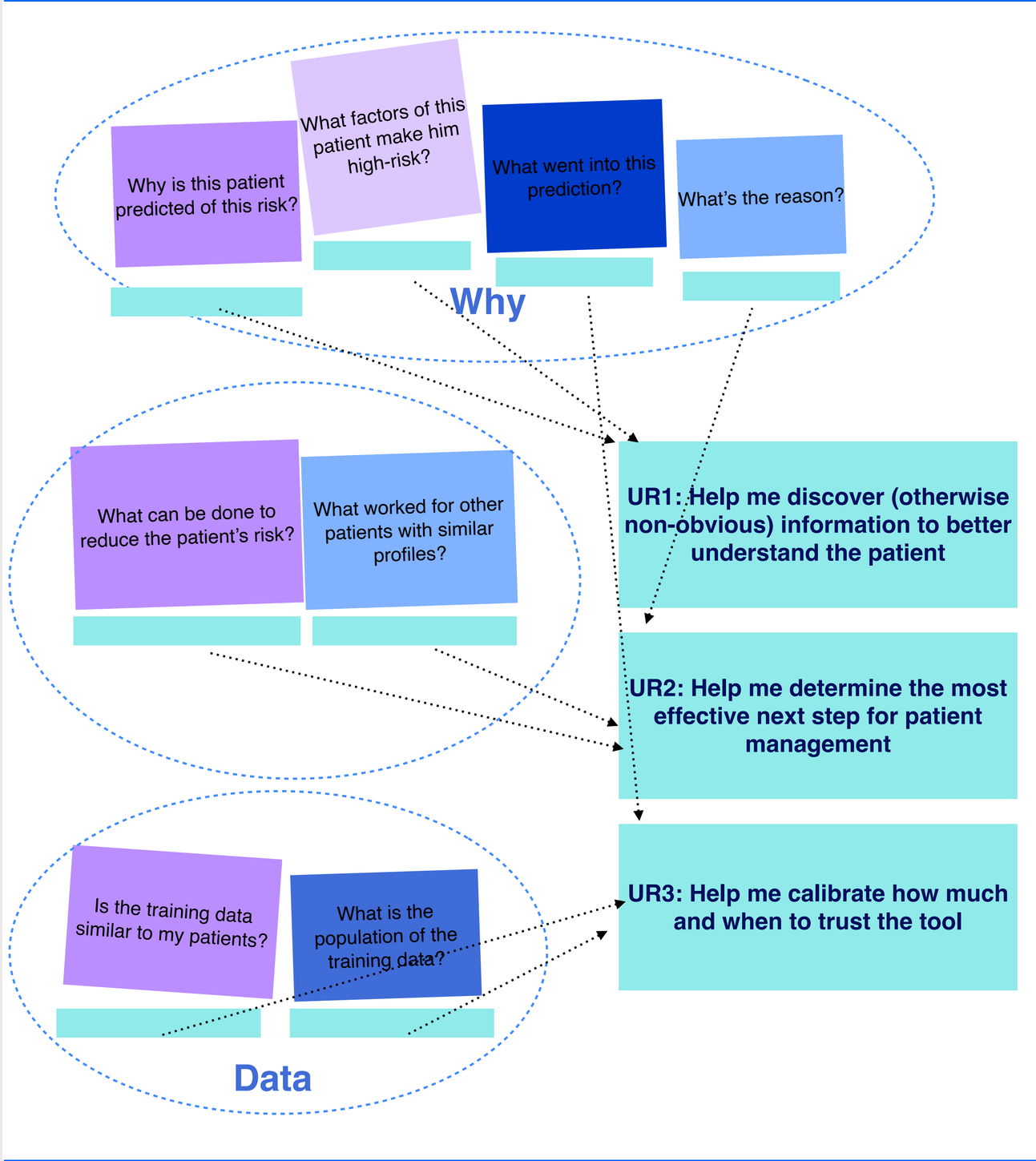
Step 2: **Analyze** to identify key user questions and user requirements

Cluster similar questions across users into key categories.

Question Bank could guide the analysis.

Identify which categories should the team focus on

Analyze user comments on intention and expectation of asking the questions to identify key user requirements



Step 3: **Map** questions to modeling solutions (with the team)

Work with data scientists and the team to map prioritized questions to explanations that the model(s) could provide

Sometimes the explanation can be derived directly from descriptive information of the model

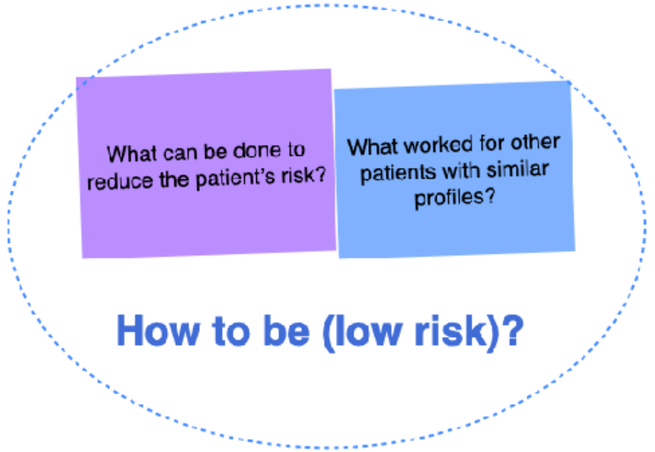
Sometimes the explanation requires implementing another set of algorithm to generate.

Step 3: **Map** questions to modeling solutions (with the team)

Work with data scientists and the team to map prioritized questions to explanations that the model(s) could provide

Sometimes the explanation can be derived directly from descriptive information of the model

Sometimes the explanation requires implementing another set of algorithm to generate.



How to be that	Highlight features that if changed (perturbed, absent, present) could change the prediction

Risk factor to eliminate	Risk improvement
UTI	-10%
Allergic reaction	-9%
Nutritional deficiency	-7%
Diabetes	-3%

Step 4: Iteratively **design** and **evaluate**

Evaluate the design, ideally with user feedback, focusing on the **key use requirements**

Identify gaps and iteratively improve the design and the modeling solution

Risk factor to eliminate	Risk improvement
UTI	-10%
Allergic reaction	-9%
Nutritional deficiency	-7%
Diabetes	-3%

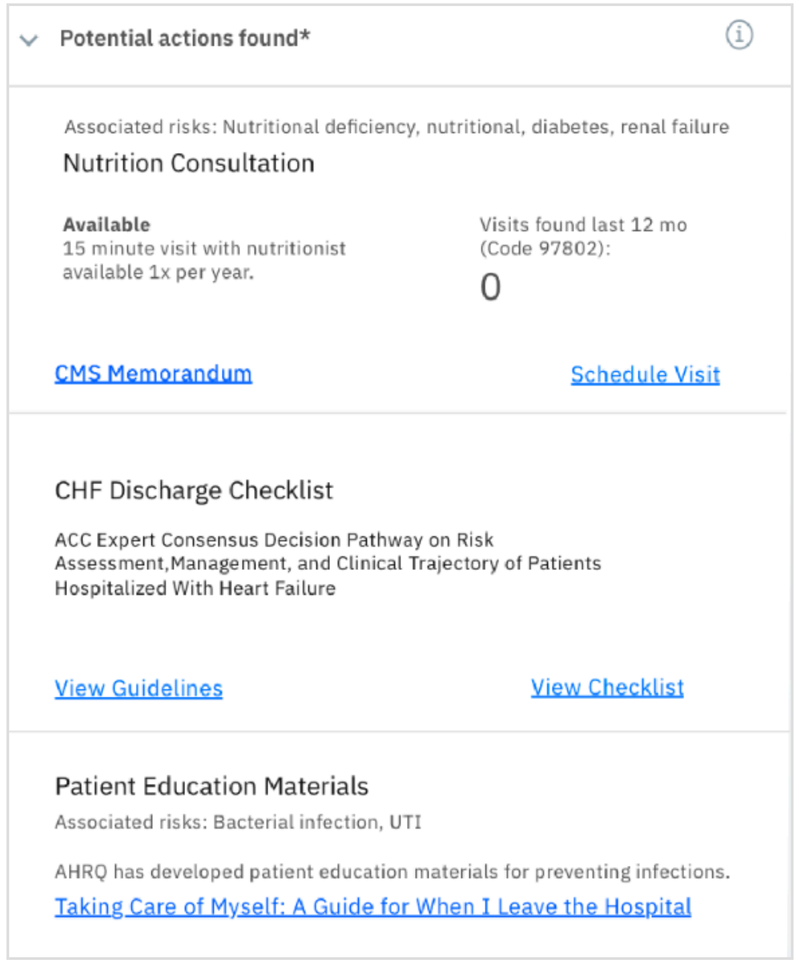
UR2: Help me determine the most effective next step for patient management

- The design lacks “actionability”. User wishes to see “how-to”
- Some factors are not easy or possible to eliminate

Step 4: Iteratively **design** and **evaluate**

Evaluate the design, ideally with user feedback, focusing on the **key use requirements**

Identify gaps and iteratively improve the design and modeling solution

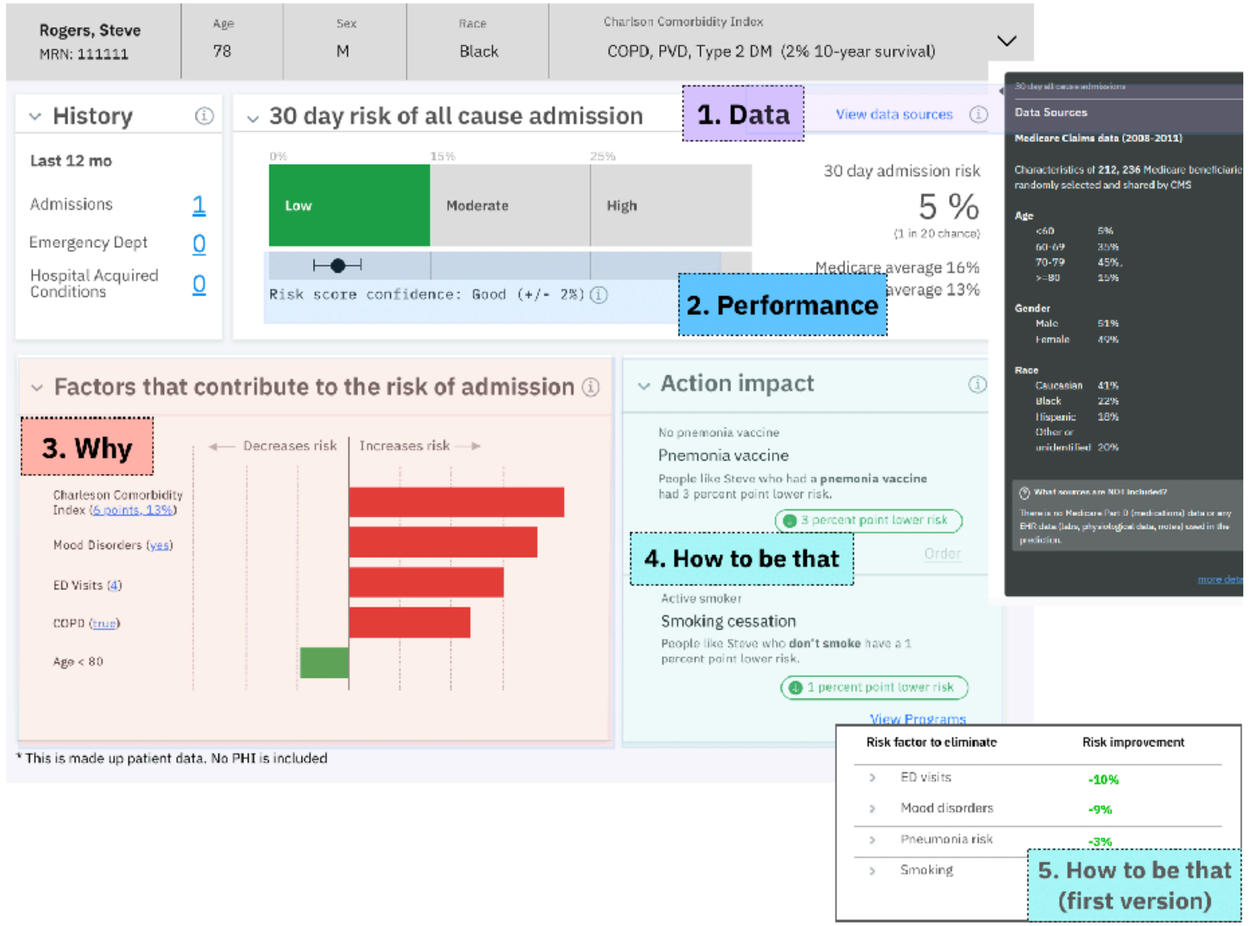


- Link to external guidelines for “how-to”
- Add modeling constraints based on the “costs” of changing risk factors

Step 4: Iteratively **design** and **evaluate**

Evaluate the design, ideally with user feedback, focusing on the **key use requirements**

Identify gaps and iteratively improve the design and modeling solution



A Few High-Level Ideas…

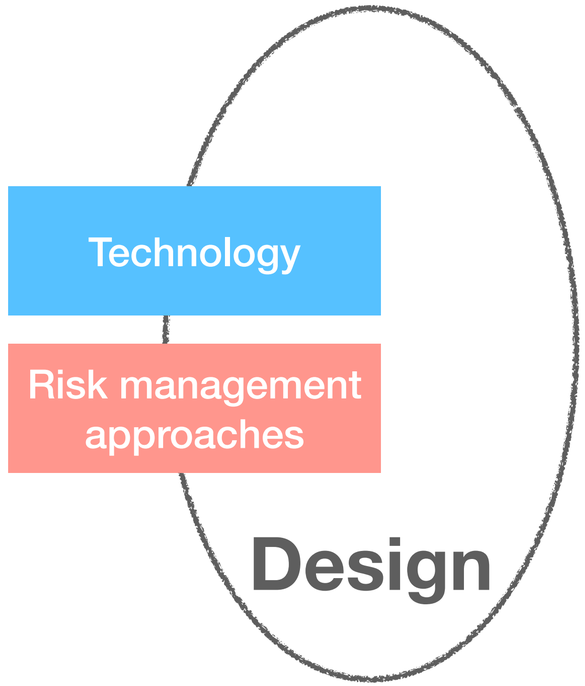
Challenge with understanding the material

Challenge with choosing (and having) the right material

Challenge with mutual shaping of design and material

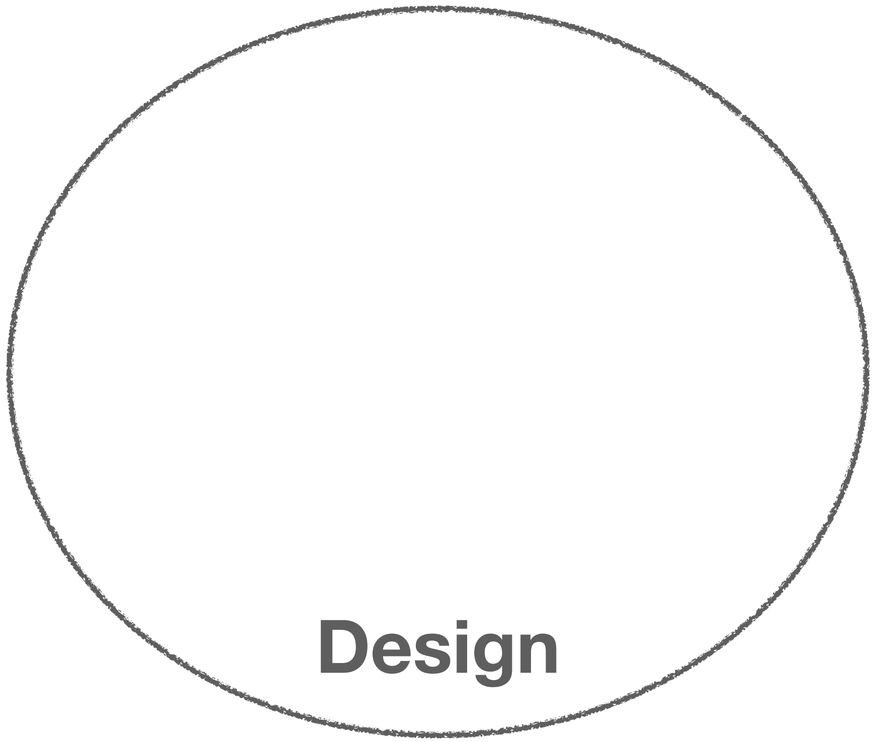
Reframing the technical space by “socio” requirements

Facilitating designer-enginer collaboration with a shared workflow and boundary objects



Technology goals

Risk assessment



Technology

Risk management approaches

Technology goals

Risk assessment

Research Thread 2: Empower Designers in the Age of GenAI

Working with “pre-trained” models powering heterogenous applications

One “foundation” model

A widening sociotechnical gap?

Many heterogenous “socio” contexts

Deep learning

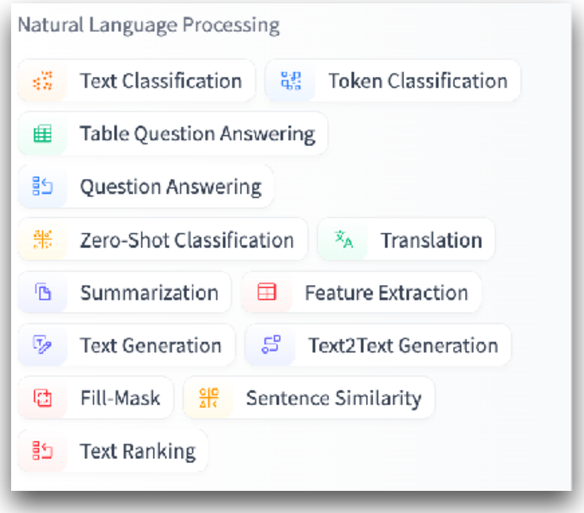
2016

Responsible AI

Pre-trained “small” models that perform specialized tasks

2021

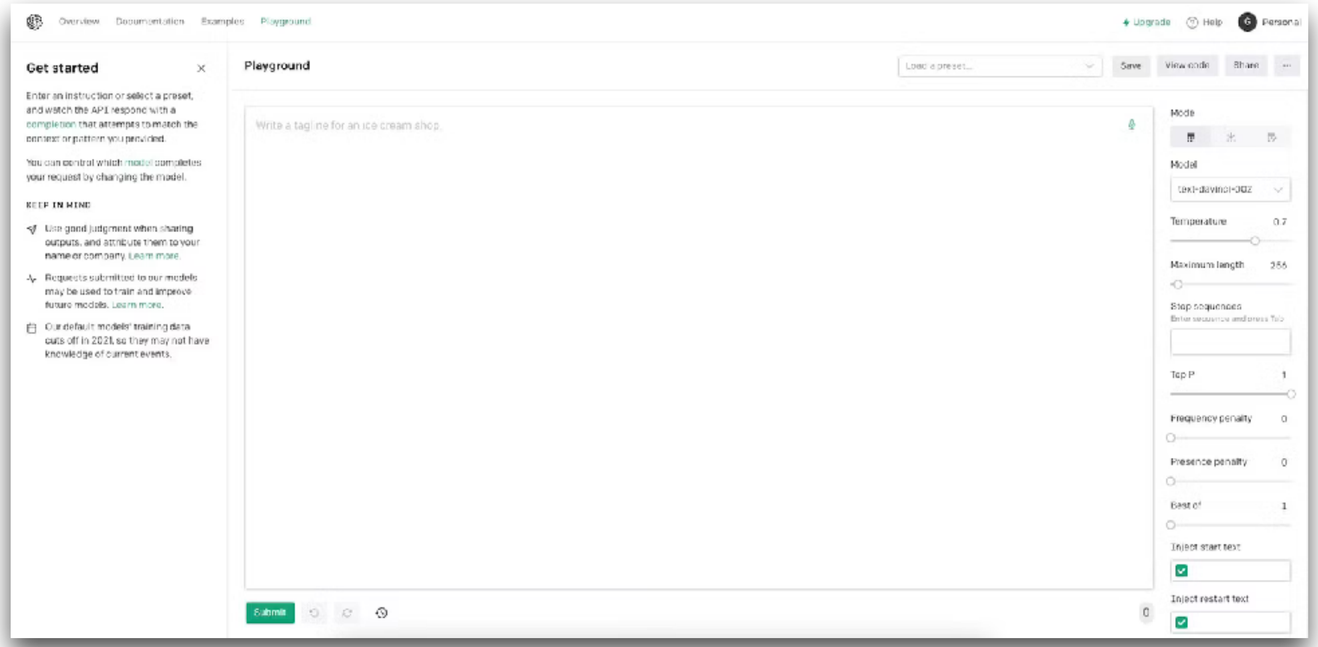
Generative AI



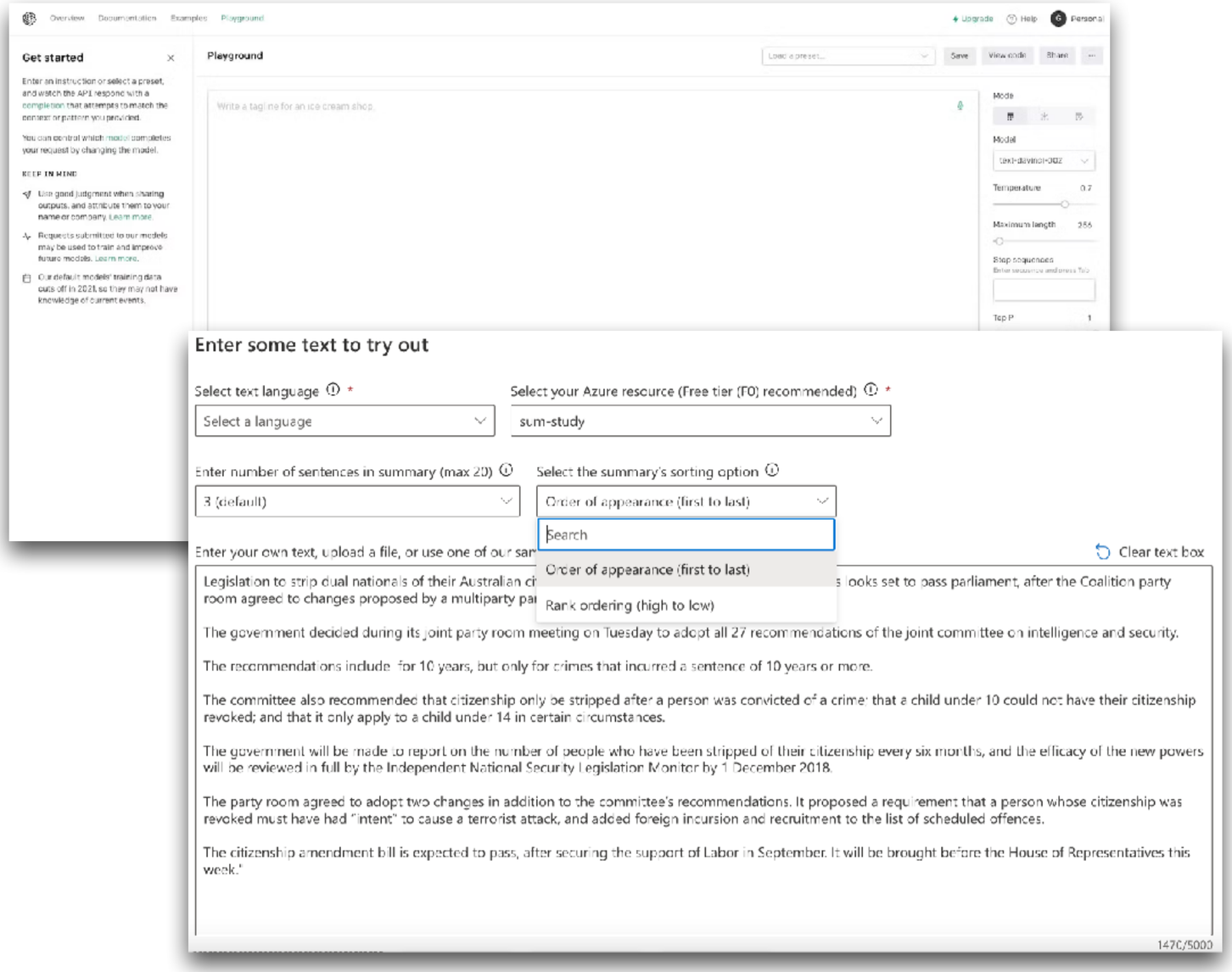
Opportunities with pre-trained models

Lower barriers for using models in product development

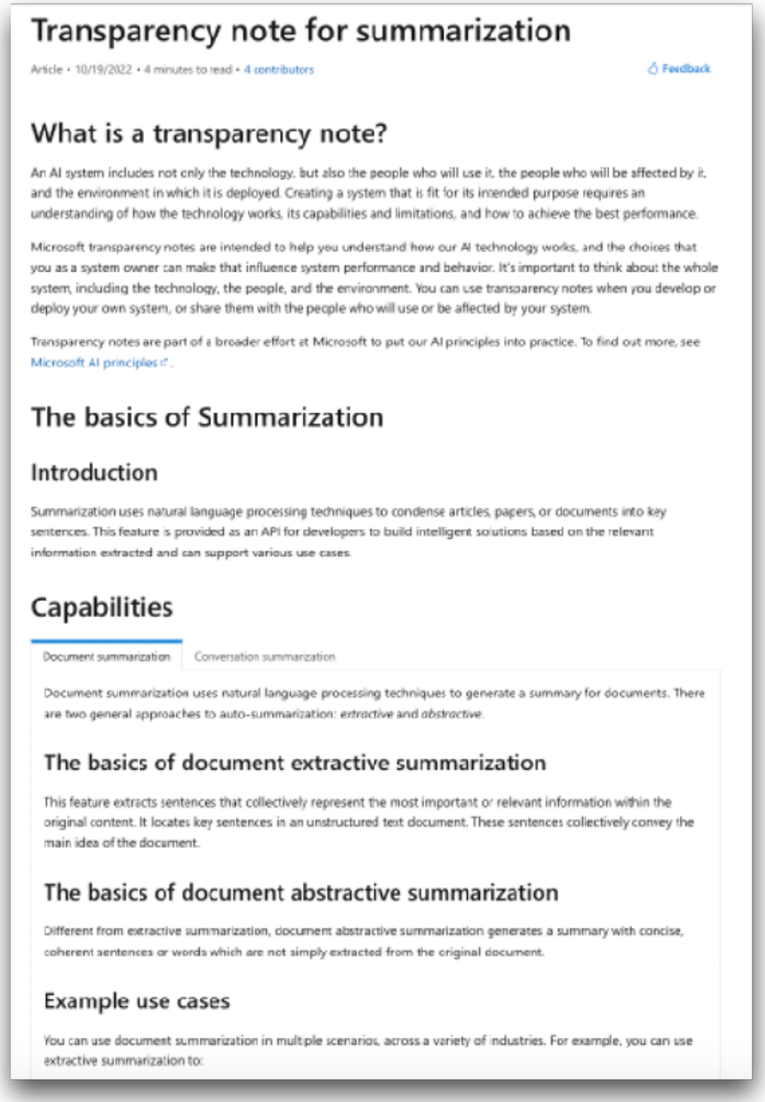
Allow designers to directly explore the design materials



Playground UI: understanding by trying out examples



Playground UI: understanding by trying out examples



Transparency note for summarization

Article • 10/19/2022 • 4 minutes to read • 4 contributors

Feedback

What is a transparency note?

An AI system includes not only the technology, but also the people who will use it, the people who will be affected by it, and the environment in which it is deployed. Creating a system that is fit for its intended purpose requires an understanding of how the technology works, its capabilities and limitations, and how to achieve the best performance.

Microsoft transparency notes are intended to help you understand how our AI technology works, and the choices that you as a system owner can make that influence system performance and behavior. It's important to think about the whole system, including the technology, the people, and the environment. You can use transparency notes when you develop or deploy your own system, or share them with the people who will use or be affected by your system.

Transparency notes are part of a broader effort at Microsoft to put our AI principles into practice. To find out more, see Microsoft AI principles.

The basics of Summarization

Introduction

Summarization uses natural language processing techniques to condense articles, papers, or documents into key sentences. This feature is provided as an API for developers to build intelligent solutions based on the relevant information extracted and can support various use cases.

Capabilities

Document summarization | Conversation summarization

Document summarization uses natural language processing techniques to generate a summary for documents. There are two general approaches to auto-summarization: *extractive* and *abstractive*.

The basics of document extractive summarization

This feature extracts sentences that collectively represent the most important or relevant information within the original content. It locates key sentences in an unstructured text document. These sentences collectively convey the main idea of the document.

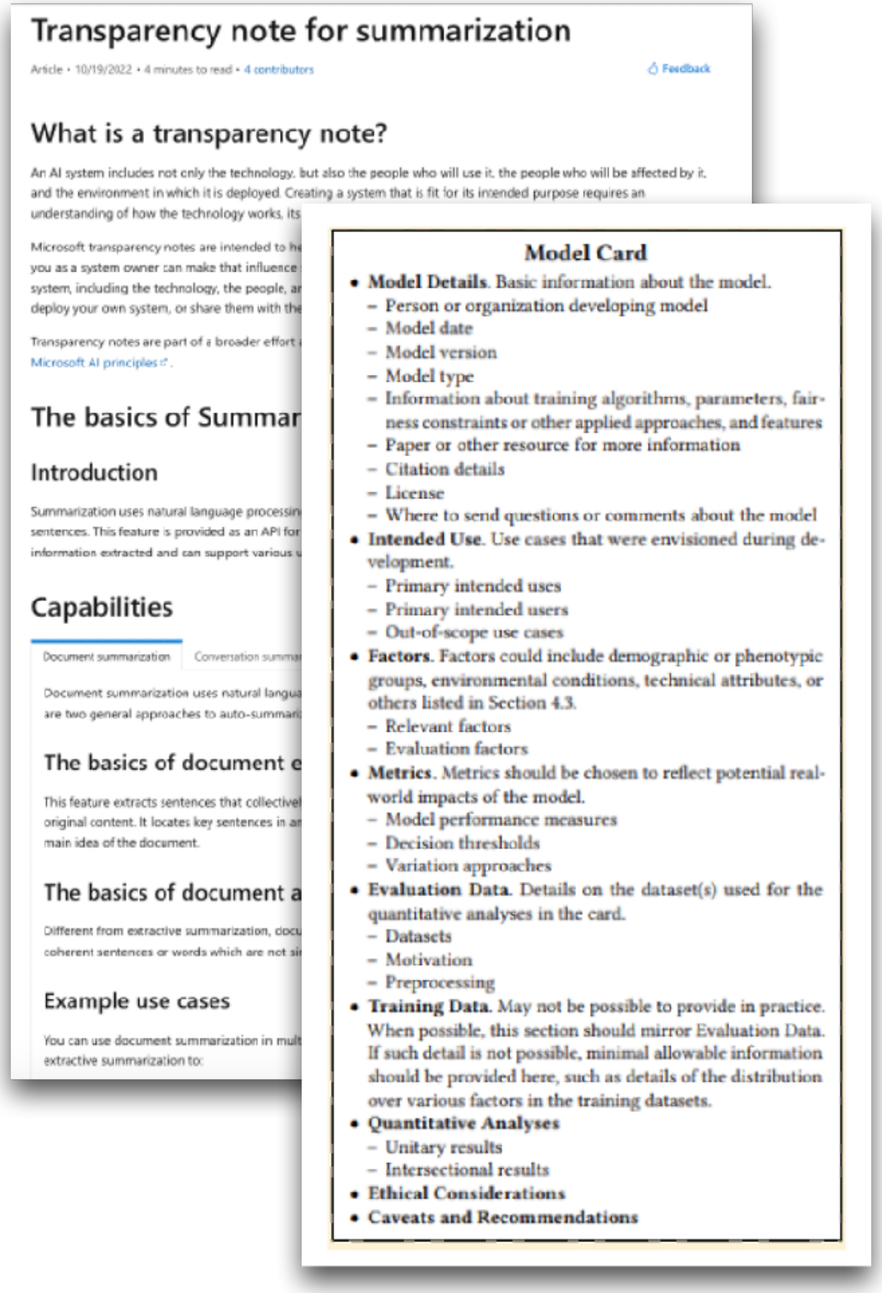
The basics of document abstractive summarization

Different from extractive summarization, document abstractive summarization generates a summary with concise, coherent sentences or words which are not simply extracted from the original document.

Example use cases

You can use document summarization in multiple scenarios, across a variety of industries. For example, you can use extractive summarization to:

Model Documentation/ Card: “nutrition label” of a model



Transparency note for summarization

Article • 10/19/2022 • 4 minutes to read • 4 contributors

Feedback

What is a transparency note?

An AI system includes not only the technology, but also the people who will use it, the people who will be affected by it, and the environment in which it is deployed. Creating a system that is fit for its intended purpose requires an understanding of how the technology works, its

Microsoft transparency notes are intended to he
you as a system owner can make that influence
system, including the technology, the people, ar
deploy your own system, or share them with the

Transparency notes are part of a broader effort
Microsoft AI principles .

The basics of Summar

Introduction

Summarization uses natural language processin
sentences. This feature is provided as an API for
information extracted and can support various u

Capabilities

Document summarization | Conversation summa

Document summarization uses natural langua
are two general approaches to auto-summari

The basics of document e

This feature extracts sentences that collectivel
original content. It locates key sentences in ar
main idea of the document.

The basics of document a

Different from extractive summarization, docu
coherent sentences or words which are not si

Example use cases

You can use document summarization in mult
extractive summarization to:

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Model Documentation/ Card: “nutrition label” of a model

Purposeful use: *Should I use a model? If so, <u>where</u>?*

What interactions is the model suitable for?

How to design the interactions?

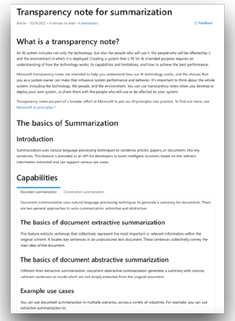
Needs Finding for “Designedly Understanding” of Pre-trained Models

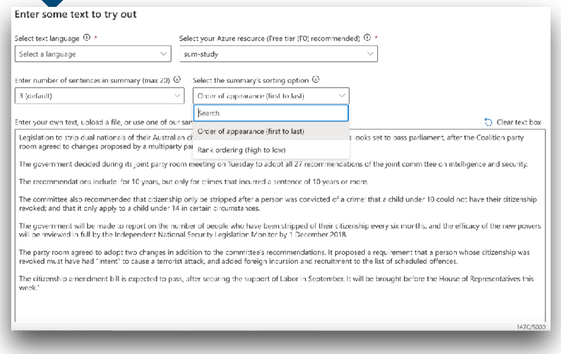
Design task to use a pre-trained model

Liao et al. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. *CHI 2023*

Needs Finding for “Designedly Understanding” of Pre-trained Models

Design task to use a pre-trained model



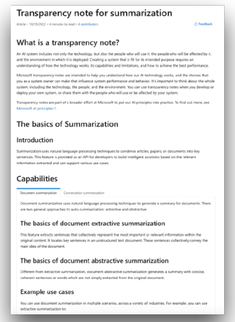


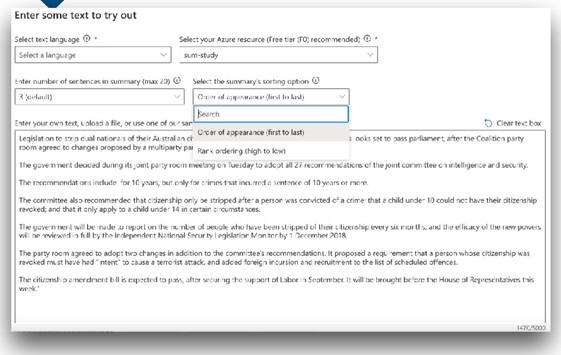
Read a model card and browse examples on playground UI

Liao et al. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. *CHI 2023*

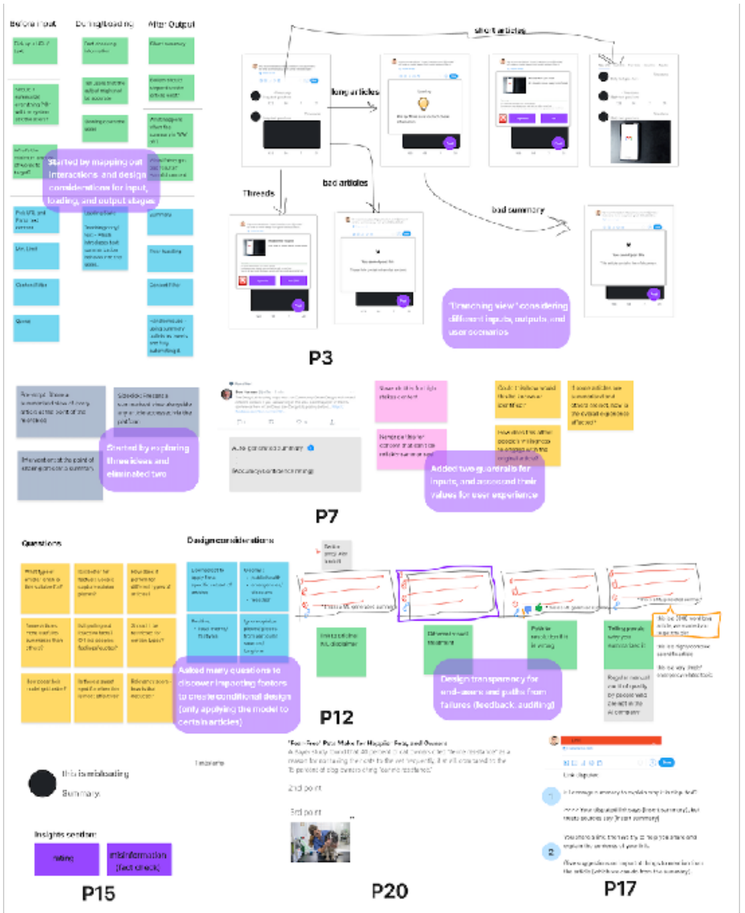
Needs Finding for “Designedly Understanding” of Pre-trained Models

Design task to use a pre-trained model





Read a model card and browse examples on playground UI



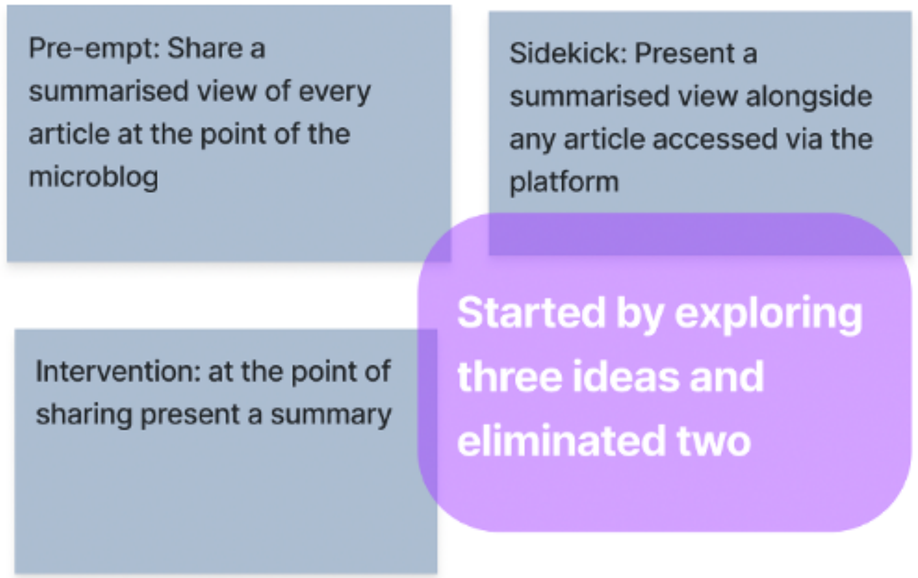
Perform design ideation

Liao et al. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. *CHI 2023*

Transparency goal	Provided info used	Requested information
G1: Divergent-convergent design thinking	intended uses, model description, input-output examples, harms considerations, unintended uses, limitations	output analysis, explanations
G2: Conditional design	impacting factors in limitations and harms considerations, examples of different categories	training data, explanations, disaggregated evaluation with performance and other output characteristics, confidence/ uncertainty
G3: Transparency for users	model description, limitations, harms considerations	performance, confidence/uncertainty, explanations
G4: Team negotiation and collaboration	harms considerations, limitations, design space guidance	customizability and improvability, algorithm, training data and other development information

Transparency goal	Provided info used	Requested information
G1: Divergent-convergent design thinking	intended uses, model description, input-output examples, harms considerations, unintended uses, limitations	output analysis, explanations
G2: Conditional design	impacting factors in limitations and harms considerations, examples of different categories	training data, explanations, disaggregated evaluation with performance and other output characteristics, confidence/ uncertainty
G3: Transparency for users	model description, limitations, harms considerations	performance, confidence/uncertainty, explanations
G4: Team negotiation and collaboration	harms considerations, limitations, design space guidance	customizability and improvability, algorithm, training data and other development information

Designers' goal 1 with understanding**: eliminating "risky" design ideas to use the model**



P3

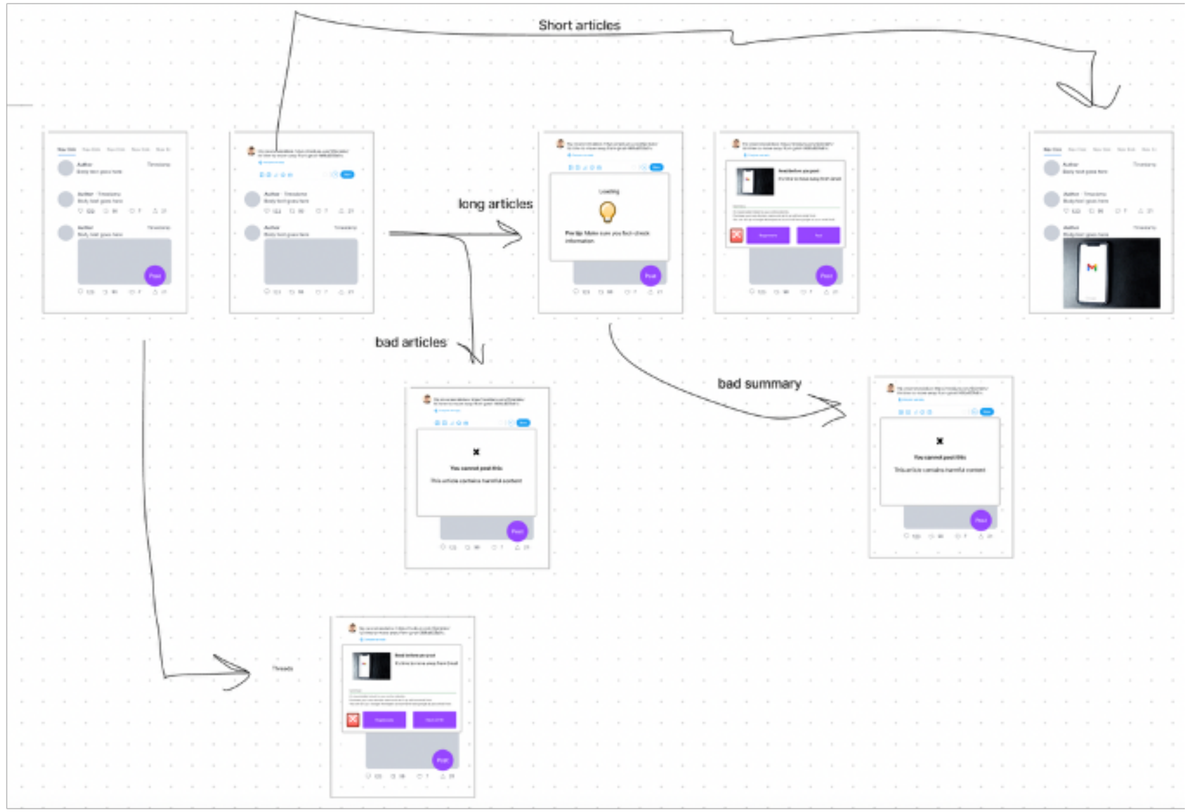
“I have the question of how reliably it could perform… if it was an intervention and it was unreliable...you’re out of your extra step and it’s literal nonsense. And that really diminishes somebody’s experience with the whole product, so that presents, I think, a huge risk.”

Require discovering model limitations situated in different designs

Formulating hypothesis by translating from sociotechnical risks, but challenging to validate with current support

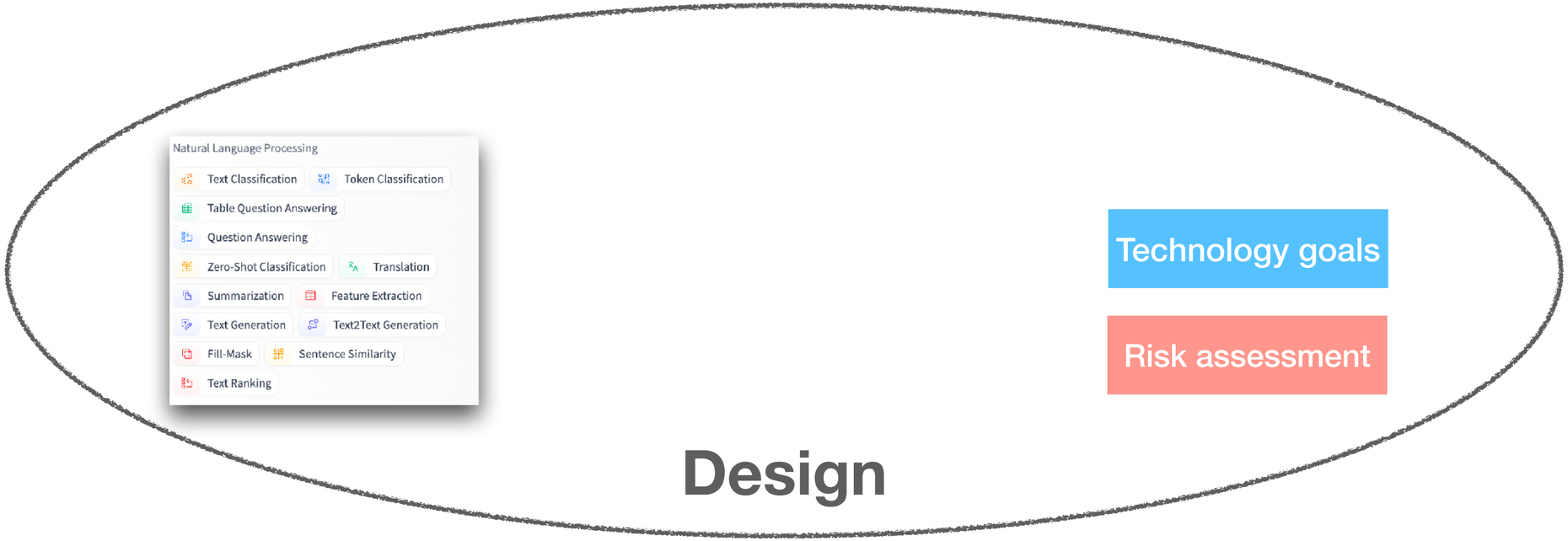
Transparency goal	Provided info used	Requested information
G1: Divergent-convergent design thinking	intended uses, model description, input-output examples, harms considerations, unintended uses, limitations	output analysis, explanations
G2: Conditional design	impacting factors in limitations and harms considerations, examples of different categories	training data, explanations, disaggregated evaluation with performance and other output characteristics, confidence/ uncertainty
G3: Transparency for users	model description, limitations, harms considerations	performance, confidence/uncertainty, explanations
G4: Team negotiation and collaboration	harms considerations, limitations, design space guidance	customizability and improvability, algorithm, training data and other development information

Designers' goal 2 with understanding**: coping with model limitations by creating "conditional design" and guardrails**



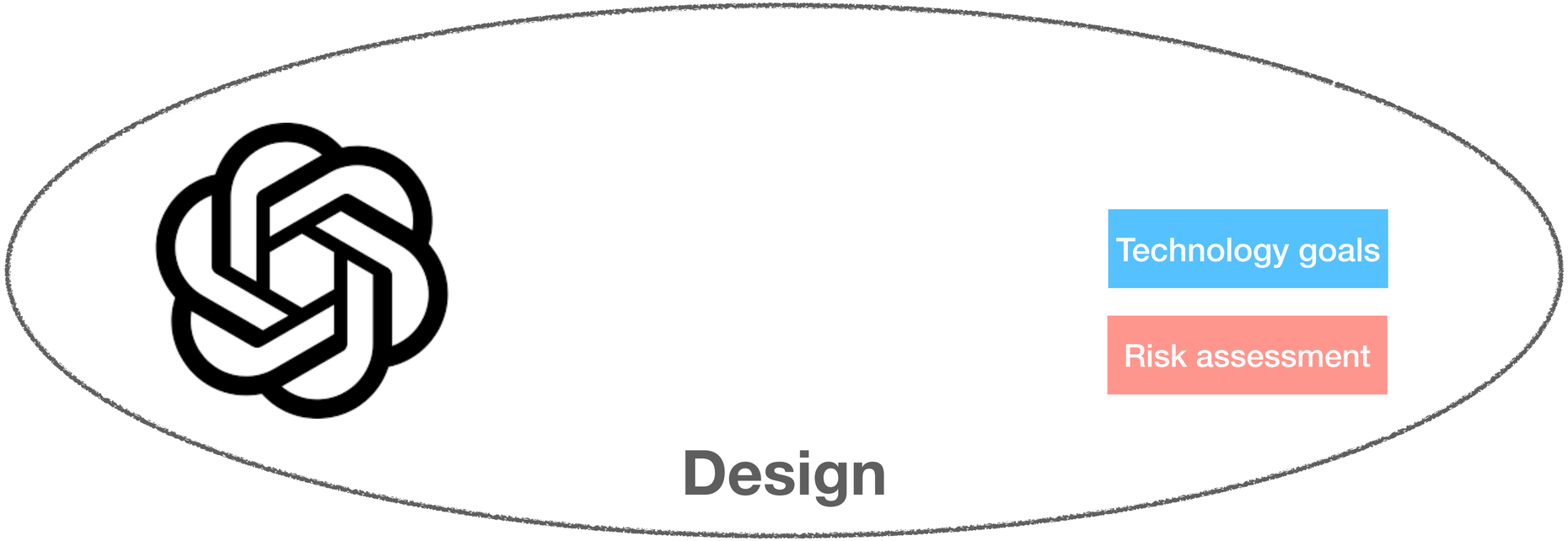
Experienced AI designers gravitate towards creating different designs or guardrails for different types of model inputs or outputs

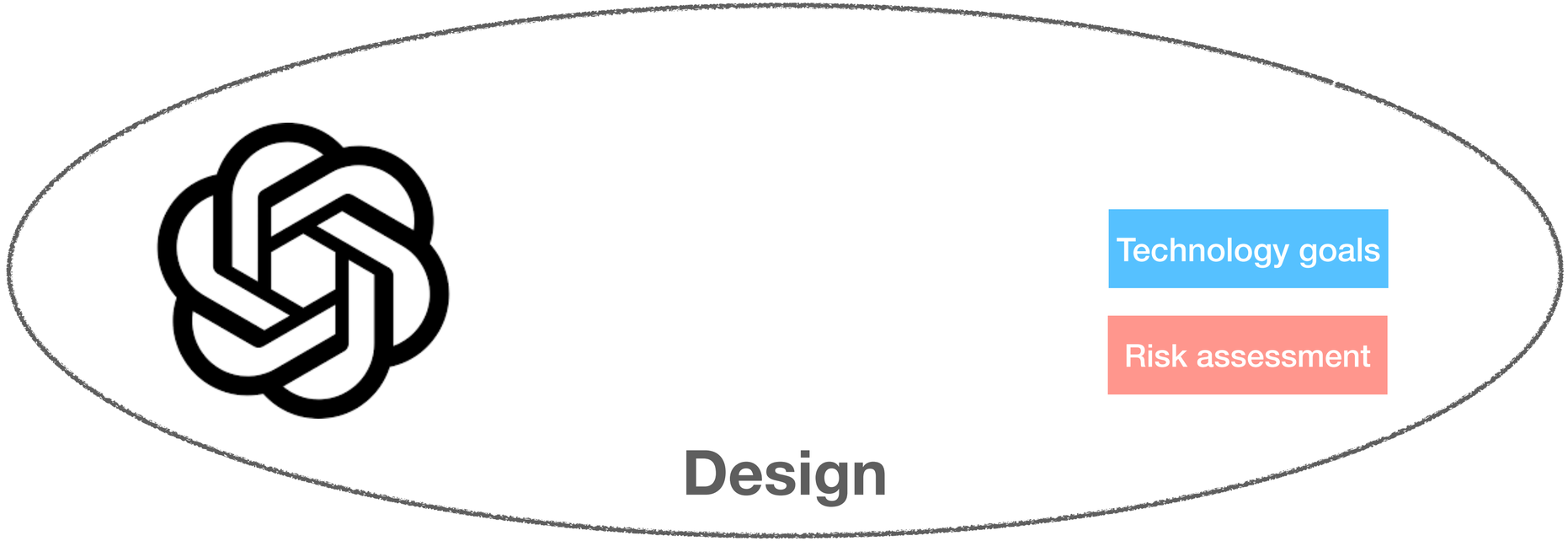
Require discovering “impacting factors” that vary model behaviors



Pre-trained models give designers more autonomy and control over the exploration of design materials and design space

Design should take a more central role as **purposeful use** and **coping with model limitations** become *the* primary tasks with responsible use of pre-trained models

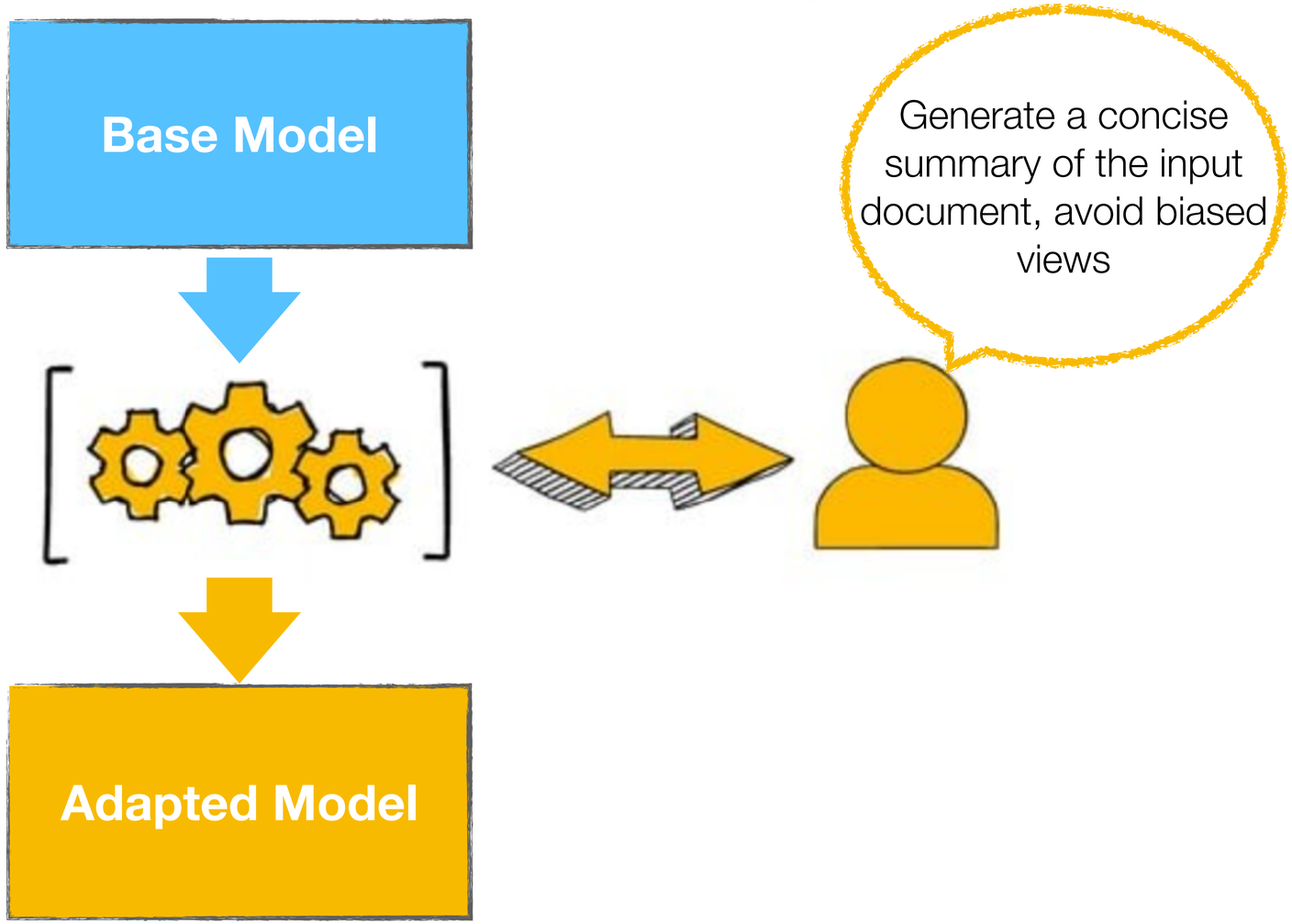

Technology goals
Risk assessment
Design



What is the unique materialistic property of current “large” GenAI models (e.g. large language models)?

Adaptability: Central Materialistic Property of GenAI

Meta-prompt/system message: a prompt that instructs the model, applied to all (or a type of) user inputs



Adaptability: Central Materialistic Property of GenAI

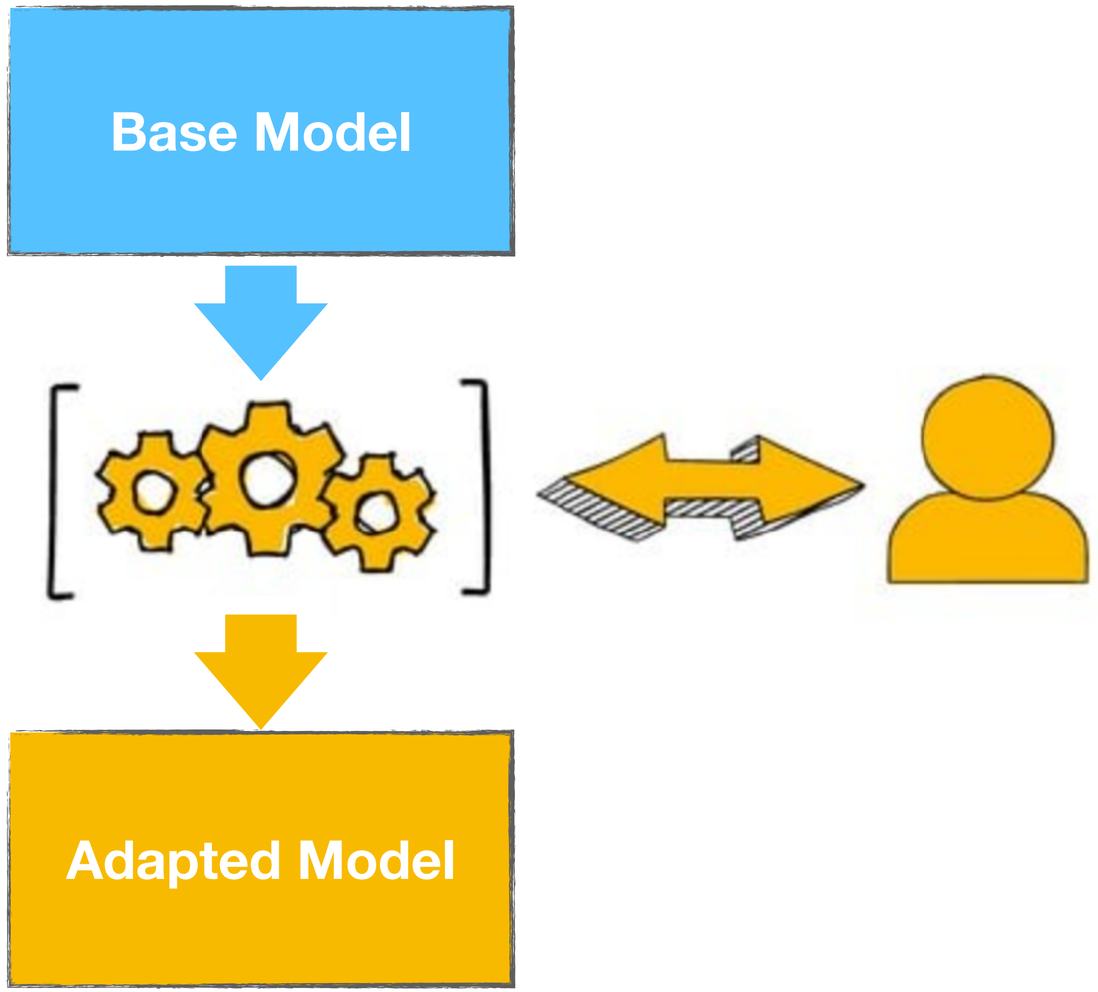
Meta-prompt/system message: a prompt that instructs the model, applied to all (or a type of) user inputs

Fine-tuning

Knowledge base grounding

Safety filters

…

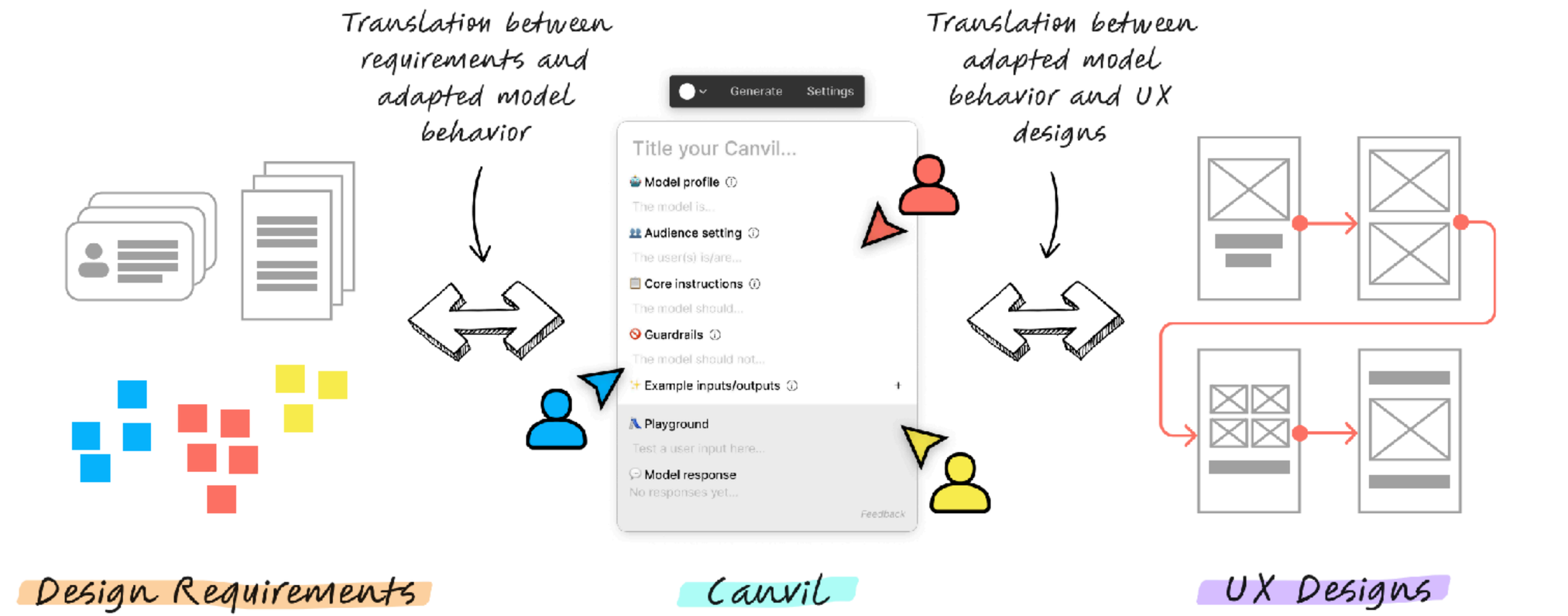


Model adaptation as tinkering with the design material

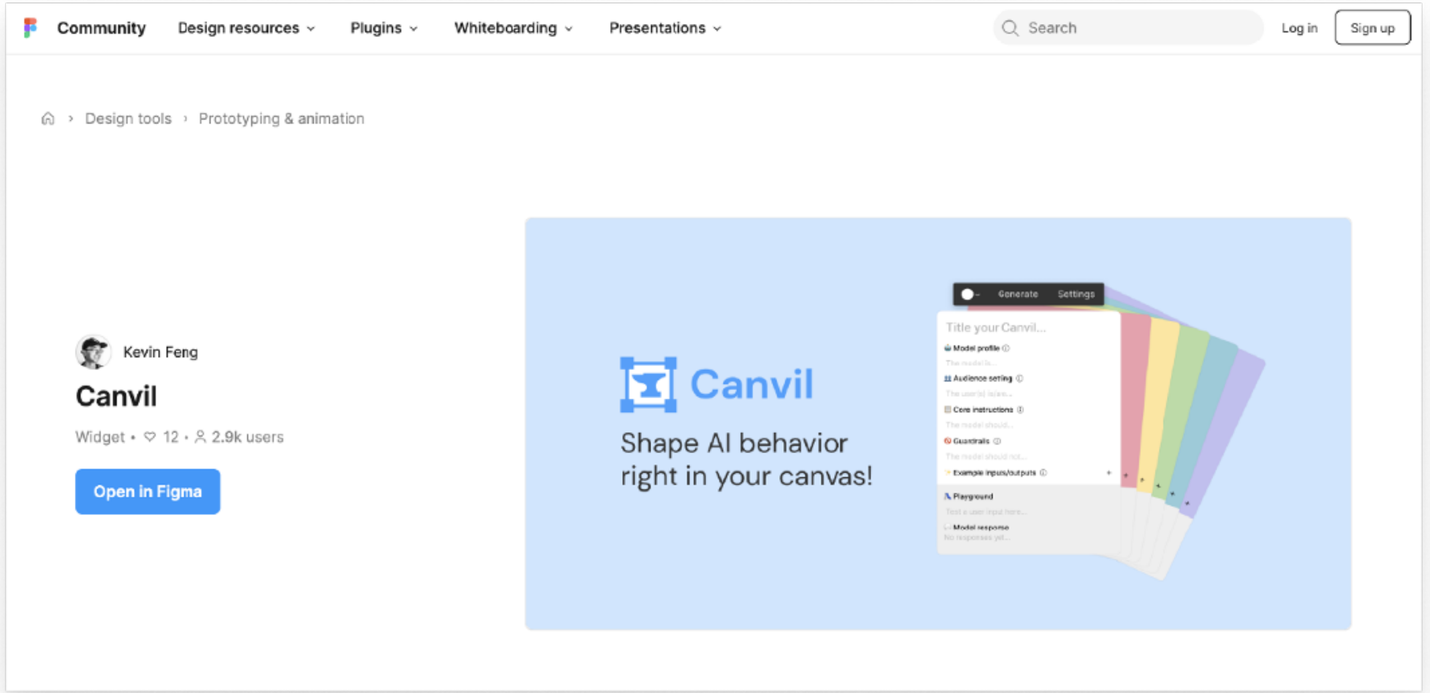
Fast, iterative mutual shaping of design and material

Encoding design requirements in meta-prompts

Canvil: A Figma Widget for Designedly Adaptation



Feng et al. Canvil: Designerly Adaptation for LLM-Powered User Experiences. *CHI 2025*



Try Canvil out!

Feng et al. Canvil: Designerly Adaptation for LLM-Powered User Experiences. *CHI 2025*

Paths Forward:

How should UX design evolve to meet the requirements of RAI?

Paths Forward:

~~How should UX design evolve to meet the requirements of RAI?~~

How can RAI better leverage UX design expertise?

Toolbox

Training

Organizational practices

Toolbox

RAI principle specific design patterns and methods

Expand UX evaluation with situated model evaluation and risk assessment

Orient design on “envisioning the sociotechnical”

Training

Organizational practices

Toolbox

RAI principle specific design patterns and methods

Expand UX evaluation with situated model evaluation and risk assessment

Orient design on “envisioning the sociotechnical”

Training

Beyond AI literacy, opportunities to explore and tinker

Strengthen and pride the critical lens

Reposition UX as a resource that cuts through AI development lifecycle (and policy)

Organizational practices

Toolbox	Training	Organizational practices
RAI principle specific design patterns and methods	Beyond AI literacy, opportunities to explore and tinker	Involve design expertise early and often
Expand UX evaluation with situated model evaluation and risk assessment	Strengthen and pride the critical lens	Rethink the “separation of concern” practice. Break expertise and cultural barriers
Orient design on “envisioning the sociotechnical”	Reposition UX as a resource that cuts through AI development lifecycle (and policy)	Define and incentivize new design roles for RAI

Thank **YOU!**

And thanks to my amazing collaborators:

Alexandra Olteanu, Alison Smith-Renner, Ameneh Shamekhi, Amit Dhurandhar, Anna Kawakami, Bart P.Knijnenburg, Bhavya Ghai, Biplav Srivastava, Bruce Schatz, Claudia Wagner, Daniel Russell, Daricia Wilkinson,, David Piorkowski, April Yi Wang, Casey Dugan, Chacha Chen, Chenhao Tan, Dakuo Wang, Daniel Gruen, Elizabeth Daly, Finale Doshi-Velez, Gagan Bansal, Gloria Mark, Hal Daume III, Hariharan Subramonyam, Javier Antorán, Jennifer Wang, Jiao Sun, Jennifer Wortman Vaughan, Jian Zhao, Jina Suh, Jonathan Dodge, Justin D. Weisz, Kartik Talamadupula, Khai Truong, Klaus Mueller, Koustuv Saha, Kush R. Varshney, Mark Riedl, Marina Danilevsky, Matthew Davis, Markus Strohmaier, Mayank Agarwal, Michelle Zhou, Michael Muller, Michael Hind, Mingming Fan, Mohit Jain, Nikola Banovic, Öznur Alkan, Peter Pirolli, Prasanna Sattigeri, Pratyush Kumar, Praveen Chandar, Rachel Bellamy, Rajan Vaish, Ronny Luss, S. Shyam Sundar, Samuel Carton, Samir Passi, Sarah Miller, Shamsi T. Iqbal, Shreya Chowdhary, Shweta Narkar, Shwetak Patel, Snehal Prabhudesai, Soya Park, Stephanie Houde, Steven Moore, Su Lin Bloddget, Sumit Asthana, Susu Zhang, Thomas Erickson, Umang Bhatt, Upon Ehsan, Valerie Chen, Victoria Bellotti, Vivian Lai, Wai-Tat Fu, Werner Geyer, Xianyou Yang , Xun Huan, Yiming Zhang, Yunfeng Zhang, Yunyao Li, Zahra Ashktorab, Ziang Xiao